TECHNICAL DOCUMENTATION

May 2022

WHAT IT IS

Origins is a system that classifies consumers according to the part of the world from which their forebears are most likely to have originated.

Each consumer on a customer file can be placed into one of 250 different origins types on the basis of their personal and family names. The segmentation could be used for example to identify people on a customer file whose ancestry is most likely to be from Ireland, Italy, or Albania.

Origins uses the same information to code customers on the basis of their most likely language and religion. An age estimate and gender can also be appended for most customers.

HOW IT'S USED

Origins is used in three different ways:

1. Origins is used to **profile customers and customer segments**. By profiling customers you can identify which groups are under or over-represented on your customer file. You can find out which groups prefer to use which products, channels and outlets, which ones you are good or poor at retaining and which are responsive to which types of promotion or reward.
2. Origins is used to **code customers**. By coding customers you can target campaigns to improve awareness and take up of public services by members of specific minority groups. You can also target products, such as cosmetics, media channels and travel, at audiences for whom they have be especially developed.
3. Origins is used to **classify postal codes**. Using a table which identifies the dominant Origins type in each postal code you can identify the locations in which individual communities have established themselves right down to street level (not available in Canada).

WHO USES ORIGINS

Origins has both government and the commercial sector uses. Government organizations that use Origins to inform the planning of their services to minority groups include Strategic Health Authorities, Primary Healthcare Trusts and individual Hospital Trusts. Origins has also been used by the police and by the campaign departments of political parties.

It has been used by retailers to identify the minority groups that use their services and to identify heavy users and users of particular products. Origins has also been used in the charity sector to identify the representativeness of members and donors.

HOW IT WORKS

In order to code individual customers, Origins makes use of a table which contains information on over 500,000 personal names and over 1.5 million family names. Each of these names has been examined in such a way as to identify the Origins type to which it is most likely to belong. This evaluation makes use of a number of criteria including the Origins codes of the surnames held by bearers of each personal name, and the appearance of diagnostic letter sequences. This evaluation also establishes the confidence with which we can say a particular name belongs to a particular Origins type.

Looking at the codes associated with both the personal name and the family name, and taking into account the confidence level of each, Origins identifies the Origins type to which each customer name is most likely to belong.

FEATURES

Origins types and groups can be appended to customer records using the Origins software application. This system is licensed to clients in Canada by Environics Analytics. The application is downloadable from the internet and makes use of files which themselves are updated on a regular basis as names from more countries are introduced. The license fee depends upon the version of the application licensed.

ACCURACY

The level of accuracy 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 Caribbean's.

As would be expected the system is more accurate when coding names to a general category, such as South Asians or Greeks, than to specific sub-categories, such as Sri Lankans or Greek Cypriots. 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 avoiding communicating with individuals who are least likely to belong to the selected target group.

EXPORTING PROCESSED DATA

The Export Data Wizard allows you to generate a file which includes the person name and family name fields on the file you submitted for coding together with the ID and with a number of new fields added.

Exactly what fields you are able to have added will depend upon:

1. The nature of the licence.
2. The country views.

SELECT A DESTINATION

The browser button allows you to activate a file description with the 'Destination' bar. This identifies the name and the location of the file you wish to export. The default is output.xls, which is saved on your desktop.

CONFIDENCE LEVEL

The value in the confidence level field indicates the extent to which each particular name on your input file can be confidently assumed to belong to the Origins type to which it has been assigned. High values will appear in the confidence level field where the two elements of the name originate from the same Origins category. You will also find high confidence level values where the names are particularly strongly associated with the Origins type that they belong to. Thus people bearing the name 'Ernest', which is a strongly English name, will tend to be given a higher value in the confidence level field that people bearing the name 'Felix' which is common in many countries.

Confidence Category indicates the 'distance' or 'closeness' between the first and last name. This is easily shown in an example. Take the names:

1. Christina Chen
2. Richard Rossi
3. Wang Hao.

Origins will output a void for number one as the first and last name do not belong to the same origins group or geographical region. The third example will have a category of 'Origins' as not only do the names belong to the same grouping but also have the same 'Origins' value. The second sample would be classified as 'Group' because the first and last names are of European decent but not the same country.

Technically this comes down to a comparison of the 'Code' values. Codes are the 'AAA' or 'DDA' values associated with all names. These codes also describe the relationship between regions. For example 'AAA' would be English and 'AAX', where 'X' is another character, might be another related region within close vicinity of England. 'AXX' is yet another code but because the two codes have a single A in common (reading from right to left) then there is a relationship. This relationship also directly influences the score as names that relate have a better more 'confident' overall score because the names are related in some fashion.

The 'Same type / Same group' indicator enables you to identify records where the personal name and the family name belong to the same Origins type; whether they belong to different Origins types but nevertheless fall within the same Origins groups; and whether they belong to quite different Origins groups. If you want to identify people who appear to have attributes of more than one cultural group, whether for selection or deselection from a communication, this indicator will prove very useful.

CONFIDENCE SCORES

As well as the ability to append an Origins code to a name, the Origins software identifies where the separate elements of a name originate from different cultures. The software provides a measure of confidence for the assignment of a classification to a person's name.

Two outputs generated in the Origins processing are the Confidence Category and the Confidence Score. This Information Sheet is designed to assist use and interpretation of the confidence values assigned to name records.

Confidence Category

The Confidence Category is designed to indicate whether the first name and the family name belong to the same, or to broadly similar Origins codes. If the two names are assigned to the same Origins CEL code then the value 'ORIGIN' is assigned.

If they do not belong to the same Origins CEL but do belong to CELs sharing the first two characters of the three character CEL code, they would be given the value ‘SUBGROUP’. An example is if the first name is Russian and the family name is Ukrainian. The first two characters of the CEL code for each name in this case is “DP”.

Where the Origins CEL of the first name and family names only match on the first character of the CEL code, then the value ‘GROUP’ is assigned. An example is if both names shared the character “I” (Islamic) as the first character of the CEL code.

If the personal and family names are assigned to quite different Origins groups then the category ‘VOID’ is assigned. The only exception to this rule applies to Origins Groups A (Anglo-Saxon) and B (Celtic) which are treated as synonymous. So, for example, a person with an English first name and a Scottish family name would generally be assigned to a ‘SUBGROUP’ category.

The Confidence Category influences the calculation of the Confidence Score. A name where the Confidence Category is ‘ORIGIN’ will generate a higher Confidence Score than where a name has a Confidence Category of ‘GROUP’.

Confidence Score

There is no definitive or simple way to decode or scale the confidence score. This is because of the large number of names in the system and the very large number of permutations of name combinations.

However, the following may assist in providing some guidance to interpretation and use.

The score is based on:

- Analysis of more than 1 billion individuals in the source data and the countries/regions/religions where those names are found
- The extent to which first names are associated with family names from different cultural backgrounds – particularly those that are from ‘contrasting’ cultural families
- The extent to which family names are associated with first names from different cultural backgrounds – particularly those that are from ‘contrasting’ cultural families
- The confidence score that is output is a function of the individual confidence scores for the first and family names

The higher the score the higher the confidence should be in the allocation of the name combination to the optimal CEL. In general, Values greater than 0.5 are good for all purposes. The highest score is around 15 and the lowest scores are below 0.1.

Analysts would do well to report on the distribution of confidence scores within a particular group of customers to understand the overall distribution and to assess the Origins coding characteristics of records belonging to different confidence score bands.

For most applications, a low confidence level is of no real consequence. The threshold for considering the exclusion of particular records should vary according to the use and application of the coding.

For **research and analytical purposes** – e.g. customer and area profiling - the confidence score should not be taken into account at all. This is because the code that has been assigned is the best possible assessment of cultural origin. Errors are acceptably low and the suppression of low confidence scores will not significantly change the insight from an individual profile. This is because many of the errors are self–compensating when used in aggregated analysis.

For **customer selections and targeted marketing**, the confidence score may be used to screen out those that increase the risk of inconsistency with campaign objectives or content. The cut-off should be set at a level that varies according to the nature of the campaign. As a guide, a threshold level should be set somewhere between 0.5 and 0.8, although this should be somewhat higher if the proposed campaign is designed for a very specific cultural segment – e.g. people who have recently migrated from India, or those people of Indian background who seem to retain strongest links with their cultural background.

Business rules may be developed to deal with cases where low confidence scores are a product of names that are less diagnostic of cultural origin. For example, where the first name is of Anglo-Celtic origin and the family name belongs to a cultural group that is the key subject of a campaign. It is common for many Chinese, southeast Asian and African people to adopt an Anglo-Celtic first name whilst retaining their family name. Depending on the campaign, it may be acceptable to override the low confidence scores.

Another case is where the first name is missing or is only recorded as an initial, thereby assigning the first name to an 'unknown' or 'error' classification. The family name classification may be considered a sufficiently strong indicator of cultural background and therefore meet the eligibility criteria for campaign selection.

A third case is where the conflicting CEL codes for first and family names belong to the same broad geographical region – e.g. Western Europe (e.g. French / English, Danish / Scottish, Spanish / Italian) or Asia (Indonesia / Malaysia, Laos /Thailand). The nature and content of the campaign may suggest that 'near enough' is 'good enough' and that near neighbours in the classification do not conflict with campaign objectives.

In summary, the key advice is to ignore the confidence score for all analytical purposes and to consider using the scores only when making selections for a highly targeted campaign. Even with campaign selections, it really depends on the nature and content of the campaign and an assessment of the 'risk' factor associated with a customer receiving an inappropriate communication.

ADDITIONAL ITEMS

In certain countries and where users have opted for the appropriate licence it is also possible to add to the set of columns on the export file both age and gender.

Both the age and the gender estimates are calculated by reference to the personal name. The age value runs from 0 to 9, where '0' indicates a personal name associated with the youngest adult age groups and where '9' indicates a personal name associated with the oldest adult age groups. The code 'Void' indicates that no valid personal name could be found for that name. This may either be because the name could not be recognised or because it took the form of an initial or a title. This value is organised in such a way that 10% of the adult population will have names in each of the values from 0 to 9.

The gender field identifies whether the name is associated with males (code: 'm') or with females (code: 'f'). Where it can be associated with both, as with Hilary or Robin, the code 'x' is given. The code 'x' is also given where the mix of genders is 'unknown'. The code 'Void' indicates that no valid personal name could be found for that name and the code 'Not Found' indicates that the name could not be recognised.

For more information visit https://www.originsinfo.eu/

For support, please contact your Environics Analytics representative or support@environicsanalytics.com

LIST OF GROUPS AND THEIR CODES

A	ANGLO-SAXON
B	CELTIC
C	HISPANIC
D	WESTERN EUROPEAN
E	EASTERN EUROPEAN
F	GREEK/GREEK CYPRIOT
G	JEWISH/ARMENIAN
H	BLACK AFRICAN / AFRO-CARIBBEAN
I	MUSLIM
J	SIKH
K	HINDU SOUTH ASIAN
L	EAST ASIAN
X	UNCLASSIFIED
Y	NOT RECOGNISED

LIST OF SUB-GROUPS AND THEIR CODES

AAA English

BAA Scottish

BAB Welsh

BBA Northern Irish

BBB Irish

CAA Spanish

CBA Filipino

CCA Portuguese or Brazilian

DAA French or Walloon

DCA Dutch or Flemish

DDA German

DHA Italian or Maltese

DZZ Scandinavian

EIA Polish

EJZ Czech or Slovak

EKA Hungarian

ELZ Baltic States

EMF Albanian

EMZ Formerly Yugoslav

ENA Bulgarian

EOA Romanian

EPZ Russian or Ukrainian

FAZ Greek or Greek Cypriot

GAA Jewish

GBA Armenian

HAA Black Caribbean

HBA Nigerian

HCA Ghanaian

HDZ Black South African

HED Ethiopian

HZZ Other Black African

IAZ North African Muslim

IBZ Somali

ICA Turkish

IFA Iranian

IHZ Pakistani

IIZ Kashmiri or Afghan

IKA Bangladeshi

IZZ Other Muslim

JAA Sikh

KAA Hindu Indian

KBA Tamil or Sri Lankan

KCA Bangladeshi Hindu

LAA Japanese

LAB Korean

LBA Mandarin Chinese

LBB Cantonese Chinese

LCA Vietnamese

LZZ Other East Asian

XXX Unclassified and Other