Richard Webber
is generally recognised as the originator of geodemographic classification, having been involved while working for Experian and CACI in the development of both Mosaic and Acorn. Since taking up a Visiting Professorship at University College London, he has become increasingly interested in the geography of naming practices and how government and commercial organisations can infer people's origins from their names.

This article covers subject matter which was recently highlighted in the national media

Keywords: origins, social marketing, ethnic marketing, surnames, personal names

Richard Webber
Centre for Advanced Spatial Analysis
UCL, 1–19 Torrington Place
London WC1E 7HB, UK
Tel: +44 (0)20 8340 3034
E-mail: richardwebber@blueyonder.co.uk

Using names to segment customers by cultural, ethnic or religious origin

Richard Webber

Received: 7 November 2006

Abstract

In advanced European economies, it is typical for some 20 per cent of the residential population to be either immigrants or descendants of recent immigrants. By no means are all of these people occupied in menial jobs. Indeed, a recent analysis of the British 'Rich List' suggests that these groups are now disproportionately found among the extremely wealthy. Although this 20 per cent of population are likely to have very distinctive consumer preferences, very few organisations have found effective means of identifying the extent of this population on their customer databases or of reaching them with targeted media or communications. This paper explains how, from detailed analysis of personal and family names, it may be possible to profile and target consumers according to their origins in a much more effective way than by including ethnicity, birthplace or religion on either customer or market research questionnaires.
Journal of Direct, Data and Digital Marketing Practice (2007) **8,** 226–242.
doi:10.1057/palgrave.dddmp.4350051

Introduction

Consumers born abroad or to recent immigrants represent a significant proportion of the British market. At the time of the 2001 census, 6.12 per cent of the UK population was born abroad, and 12 per cent identified themselves with ethnic groups different from White British (ethnic minorities). In London, these two population groups account for 27 and 40 per cent of the population, respectively. These figures are likely to have increased since 2001. The 2006 Pupil Level Annual School Census (PLASC) undertaken by the Department for Education and Science (DfES) suggests that as many as 17.4 per cent of children in English state schools aged 8 are now not of White British ethnicity. In Inner London the figure is as high as 71.4 per cent.

The notion that this population consists mostly of poorly educated economic migrants from developing countries, employed if at all in menial jobs, is no longer appropriate. Recent arrivals and their descendants are particularly concentrated in Greater London, the British region with the highest household incomes and wealth. Within London over 30.9 per cent of the population of the richest borough, the Royal Borough of Kensington and Chelsea, were born abroad,[1] 49.9 per cent

described their ethnicity as other than British and 73.4 per cent of 8 year olds are not of White British ethnicity. Recent analysis by the *Sunday Times*[2] of the 1,000 names on their 'UK Rich List' showed disproportionate numbers of non-British names and that a similar pattern applied among senior positions in the legal and medical professions. Meanwhile, the extent to which the 'City' now recuits its top earners from overseas is illustrated by the finding that over two-thirds of the new partners Goldman Sachs appointed in 2006 have non-British surnames.[3] Top entertainers, whether on the football pitch, the music hall or at Covent Garden, increasingly originate from overseas.

Further evidence for this proposition is that people of Jewish religion were in 2001 over ten times more likely to live in the three most affluent of Experian's Mosaic neighbourhoods, 'Global Connections', 'Cultural Leadership' and 'Corporate Chieftains', than would be expected on a random basis.[4] Analysis of the postcodes of people with Armenian names shows them living in even more upmarket neighbourhoods than the Jews.[5] An increasingly important feature of prestigious neighbourhoods in London is the number of people of Indian and Chinese origin as well as residents from Continental Europe.[6]

Rapid growth in the number of prosperous immigrants

Given the size and overall levels of consumption of the 'foreign' consumer and the rapid growth in the number of prosperous immigrants, it appears at first sight strange that this market should have been so less researched or segmented by commercial organisations than by public bodies. Agents of the state, such as schools, hospitals and the police, even councils consulting residents on the introduction of controlled parking zones, routinely include ethnicity questions on questionnaires. Not only do government departments and local authorities routinely evaluate the use of public services by minority groups, but increasingly they use information on the ethnicity of citizens to target relevant messages to them.

Ethnicity used for analysis and targeting

This practice is particularly well advanced in public health promotion. Successful recent examples of the use of ethnicity for analysis and targeting are campaigns to encourage South Asians to attend diabetes screening centres in Slough[7] and to dissuade people of Bangladeshi origin from attending accident and emergency departments of a hospital for treatments that could be adequately undertaken by general practitioners. In some of these instances the communications programmes have been driven by self-reported ethnicity data, in other cases by the use of personal and family names.

In the case of the private sector, ethnicity is more commonly asked of employees rather than customers and then often for the purpose of compliance with equal opportunities legislation. None of the principal market research surveys includes ethnicity as a routine survey code nor will ethnicity be found on the CRM systems of banks, retailers or utilities. Clearly, this is not because the topic is not relevant to consumer segmentation — asking the question and acting on the answer is perceived to be inconsistent with the 'colour blindness'

that is felt to be a politically correct feature of the behaviour of a major corporate enterprise.

Use of names to infer people's origins

The use of names to infer people's origins provides a much easier opportunity for commercial organisations to test whether their products are meeting the requirements of all segments of the market. It is also of use in the public sector either where the citizen's origin has not been stated or where data protection regulations preclude its use other than in the application for which it has been solicited. It also makes it possible to establish the manner in which different population groups like to undertake transactions, to locate retail outlets whose merchandise might appropriately be adapted to local cultural preferences and to be able to target products and services developed to meet the needs of specific population groups to individual customers who match their profile.

Examples of commercially marketed products and services that are especially likely to appeal to particular minority populations include:

— cosmetics and skin products developed to meet the needs of specific racial groupings
— foods stocked in dedicated aisles of supermarkets
— airline flights to specific destinations
— methods of recruitment for charitable causes
— television channels broadcast in particular languages
— Sharia compliant savings products
— marques of imported cars associated with country of manufacture
— culturally specific types of apparel.

Names as a basis for defining people's origins

In the late 1970s, when postcodes had recently been introduced by the Post Office, it became apparent to the marketing industry that elements of people's names and addresses could provide useful insights into their demographic profile.[8]

Elements of people's names and addresses could elucidate their demographic profile

The most intensively used of these elements is the postcode so that today there are few large consumer-facing organisations that do not generate some useful insight into their customers by appending geodemographic classifications such as Mosaic or Acorn to their marketing databases. Even before the launch of Acorn in 1979, some of the UK's largest direct marketers, such as Great Universal Stores, had sought improvements to the effectiveness of their mailing programmes by selecting or deselecting households according to elements of their address, such as the appearance of the text string 'Flat' or 'Farm'.

Subsequent to the launch of geodemographic classifications, analysts began to recognise the opportunity to make predictive inferences of people's age from their names on the basis that the fashionability of names such as 'Ivy' and 'Bunty' had long given way to names such as 'Lucinda' and 'Michelle'. Such ideas were incorporated into commercial products such as Experian's 'Stage' and CACI's 'Monica' classifications. More recently, readers of *The Times* and *The Daily*

Telegraph have been presented with statistics showing the growth and decline in fashion among different street names and house names.[9] Figure 1 illustrates in dramatic form the sum of inferences that can reasonably be made about an individual on the basis of all elements of his or her name and address.

Using names to make predictive inferences about people's ethnic or racial origins is a logical development of this process, and perhaps the only surprising aspect of this line of research is that it has not been taken seriously before. People from every culture have a name and, in almost every culture, have a personal and a family name. As a person's name is perhaps the most common item of information on a customer database and as the names that are given to children are closely rooted in the language, religion and geographical origins of their forebears, it would be surprising if the names on a database could not provide some useful information about the origins of their bearers just as customers' postcodes provide clues to their social standing and their personal names provide clues to their gender as well as life stage.

Difficulty in inferring ethnicity

In practical terms, the difficulty in inferring ethnicity is that the world's population bears so many different names, perhaps in the region of 1,000,000 recognisably different surnames and over 400,000 distinct personal names. Given that many are very local and of low frequency, how is one able to establish the origins of each individual one? To put this in perspective, the compilers of the (not so compact) Oxford Compact English Dictionary had to deal with fewer than 200,000 entries.

The most obvious strategy for compiling such a register is to base it on personal and other expert knowledge. The logic of this approach is that while the world's population may have many different names, as many as 80 per cent of people in Britain can be shown to share as few as 10,000 different family names and 320 different personal names. The application of this '80:20' rule is equally apparent in other

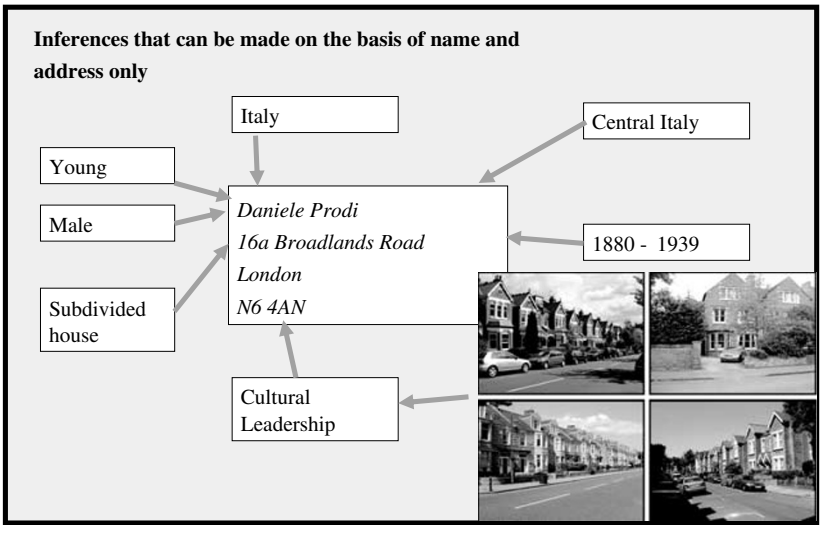

Figure 1: Inferences that can be made on the basis of name and address only

European countries. On the other hand, although personal knowledge will be effective in classifying many people, this knowledge tends by definition to be more effective in classifying people from the host population than from minorities. Unfortunately, the population groups that marketers are most interested in have names that are poorly recognised in the host community and, with a few exceptions such as 'Patel' and 'Singh', are predominately ones with low frequencies in countries other than their own homelands.

To address this problem, two of the most widely used systems for ethnicity coding in the UK, Nam Pehchan[10] and Sangra,[11] deliberately set out to tap the expert knowledge of the South Asian community regarding the origins of the names typically originating from different geographical regions, religions and linguistic groups within the sub-continent. Interestingly, both systems originated in the government sector. Nam Pehchan was created by Bradford Borough Council to assist with the city's communications with its Asian residents. Sangra, the South Asian Names Group Recognition Algorithm, was developed by the London School of Health and Tropical Hygiene with the help of specialists in public health. Both are now in common use among various health service organisations for analysing the experience of South Asians' health.

Limitations of the use of expert knowledge

The chief limitations of the use of expert knowledge are that it is time and place specific. The systems only work for the minorities for which they were built; in the case of Nam Pehchan and Sangra, South Asians and their name dictionaries only include between 3,000 and 9,000 names. Thus Nam Pehchan, which works very effectively among Asian communities that migrated to Bradford, from whose data it was built, works less well in London where names originate from other parts of South Asia and, one would suppose, would work much less well if and when applied in Germany or the USA. It is easy to overlook in this context how localised many cross-continental migration flows are — many link very specific origins and destinations. This is particularly important where a name, as is often the case, is specific to quite a localised area in the originating country.

Data mining

The alternative strategy for creating such a reference file is to apply computer-based data mining algorithms rather than expert knowledge. One example of a data mining approach, developed and used by IBM, is 'text string' analysis. This would reveal, for example, that the text string '….strom' is indicative of Swedish origin while names ending in '….porn' originate mostly in Thailand. Text strings diagnostic of other communities are shown in Table 1. Although this approach is useful for some names, the majority of names from any one country are very difficult to identify using text string search routines alone.

Another data mining approach is to compare the relative frequency of different names in different countries based on analysis of 'universal' files such as telephone directories, electoral registers, tax files and so on. Such a method would be useful, for example, in identifying 'Antonio' as both a Spanish and an Italian name but that it was marginally more common in Spain than in Italy.

Table 1: Text strings indicative of country or region of origins

Text string	Example	Likely origin
…son	Watson	England
…ie	Fairlie	Scotland
O′…	O′Sullivan	Ireland
…sma	Boersma	Netherlands
…burger	Regensburger	Germany
…dahl	Lindahl	Sweden
…es	Fernandes	Portugal
…ez	Fernandez	Spain
…elli	Martinelli	Italy
…ides	Economides	Greece / Greek Cyprus
…oglu	Demiroglu	Turkey
…ian	Aphrahamian	Armenia
El…	El Mahmoud	Middle East
…singh	Kpur-Singh	India — Sikh
…nathan	Swaminathan	India — Hindu
Ade…	Adebayo	Nigeria

It should also be noted that this addresses what can become a serious problem with systems based on expert knowledge in that the experts can fail to recognise that many names familiar among the host population, such as ‘Gill’, ‘Butt’ and ‘Lee’, can often have alternative derivations among minorities, as in these cases the Sikh, Hindu and Chinese communities. Likewise, the name ‘Tudor’ is common in both Wales and the Balkans.

Cross analysis of personal and family names

An effective data mining strategy that was pioneered by the team assembled by its editor Patrick Hanks for the creation of Oxford University Press’s Dictionary of American Family Names (DAFN)[12] involves the cross analysis of personal and family names. To build the 70,000 entry dictionary, Hanks needed to identify from among the team of national/linguistic experts on whose services he could draw which was the one to whom the family names that Hanks was not personally familiar with should be referred for etymological analysis. To facilitate this decision, the editor’s data processing manager, Ken Tucker, first obtained from Hanks a number of relatively common personal names that were particularly diagnostic of particular cultures. For example, the name ‘Brendan’ was deemed to be specific to Ireland, ‘Kurt’ to Germany, ‘Ulf’ to Sweden and ‘Mikhail’ to Russia. This information was based on expert knowledge.

Using a computerised version of the US telephone directory, Tucker next identified the frequency of each family name, in order to assist its eligibility for inclusion in the dictionary, and then identified the proportion of holders of that name whose personal names had been associated with each country. In this way, the name ‘Muller’ could be recognised as being of German origin, because its holders had far higher proportions of personal names such as ‘Kurt’, ‘Heinz’, ‘Jurgen’, etc than would be expected on a random basis, and could thence be referred to the specialist on German names.[13]

An example of how this method could have been used to identify the origin of the name ‘Lorcan’ is shown in Table 2. Taking data in this

Table 2: UK distribution of the name 'Lorcan' by 'Origin' code of surname

	English family name	Welsh family name	Scottish family name	Irish family name	Jewish family name
(a) Lorcan	30%	8%	18%	41%	2%
(b) All UK personal names	69%	11%	10%	7%	2%
(c) Ratio of (a) to (b)	0.43	0.74	1.75	5.94	1.2

case from Experian's UK Consumer Dynamics Database, we can see that as many as 40 per cent of persons bearing the name 'Lorcan' have family names originating from Ireland. This compares with only 7 per cent of all persons on the UK Consumer Dynamics Database having family names originating from Ireland. Given the frequency of occurrences of 'Lorcan' on the file, it is evident that the name 'Lorcan' must be especially associated with Ireland and hence can be assigned as Irish in any coding system. This method, which is known as 'Cultural, Ethnic and Linguistic Grouping (CELG)', can be viewed as a sort of triage and is particularly effective for identifying the correct assignment for less frequently occurring names, which is why it was used in the DAFN project.

The same general method can also be used vice versa, that is to infer the origins of personal names using the origins of the surname. A particular benefit of this approach is that it provides quantifiable evidence that can indicate the confidence with which a name can be associated with a culture. Thus, we can measure the extent to which a name such as 'Roger', whose bearer includes a Swiss tennis champion as well as an Anglo Saxon film maker, is less diagnostic of national origin than the name 'Arpad', which appears to be borne only by people of Hungarian origin.

Hitherto most projects involving the creation of name to ethnicity reference tables, such as Nam Pehchan, Sangra and DAFN, have been based on analysis of data from single countries, typically the US and UK. The approaches used by Origins, described below, and by IBM by contrast benefit from access to universal files from many countries; in the case of the Origins classification the reference file has been created using universal files from ten countries: Australia, France, Ireland, Italy, Netherlands, Norway, Romania, Spain, Sweden and the UK. Having access to files from so many countries makes it possible to improve the accuracy of the classification by incorporating into the methodology analysis of relative frequencies of names in different countries.[14]

Having access to files from several countries improves the accuracy of the classification

Using data from multiple countries also improves the reference file by identifying with greater accuracy the names specific to each country and increases the number of names covered. The system is also likely to be robust enough to be applicable to customer files from more different countries.

Assigning Origins codes to individual consumers

Database

The Origins classification applied each of the various techniques described in the previous section to the master files from the ten countries for which data could be obtained. This resulted in a database of 600,000 family names and 200,000 personal names. For the purpose of coding customer files, these reference files are loaded into software designed to enable consumer-facing organisations to optimally ethnicity code their customer name files. Although the majority of the names originate from the ten countries to which these master files relate, it can be seen from Table 3 that around a quarter of all the surnames, some 150,000, originate from parts of the world other than Europe and therefore represent the names of migrants from those countries to Europe or Australia rather than the indigenous population.

200 different Origins types

These 800,000 names were, by use of these various techniques, coded to one of 200 different Origins types, classable into the 13 principal 'Origins' groups listed in Table 3. In many cases, the Origins types are co-terminous with political boundaries. 'Hungary', 'Lithuania' and 'Myanmar' (Burma) are examples of classification codes that correspond to particular nations. Elsewhere we find instances of individual countries supporting more than one 'onomastic' category, as for example in Spain where the Basque, Galician and Catalan regions have quite distinct naming practices. In India by contrast, the basis of the distinctive subdivisions is religion (Hindu, Sikh, Muslim) and in South Africa race (Afrikaans, Black). Elsewhere we find instances where more than one country shares the same code. Germany and Austria, for example, share the code 'German' while the whole of Spanish-speaking South and Central America shares a common code because it shares broadly similar naming practices.

About five per cent of the family names in the reference file can be recognised but cannot be classified because it is impossible to find a clear identity for them.

Among the 200,000 personal names, there are a number of the more common names, such as Michael, Peter, Roger, Felix, that are common

Table 3: Distribution of family names by origin

Origins of family name	Number of names	% of names
Anglo Saxon	114,763	18.87
Celtic	34,222	5.63
Hispanic	24,072	3.96
European	280,031	46.04
Nordic	61,892	10.18
Greek Orthodox	16,007	2.63
Jewish / Armenian	3,006	0.49
African	9,184	1.51
Muslim	30,621	5.03
Sikh	3,458	0.57
Hindu	11,310	1.86
Japanese	1,525	0.25
Chinese	2,850	0.47
Origin not known	15,289	2.51
Total	608,230	100.00

to many different cultures. These 'international' names need to be considered separately both to minimise their involvement in the coding system and to prevent customers with these names being allocated the same origins code irrespective of where the files themselves originate from.

Useful to establish the origins code of both the personal and family name

To best infer the origins of a name on a customer file, it is therefore useful to establish the origins code of both the personal and the family name. In most cases the origins of the two names will be the same (such as 'Richard' and 'Webber' both being English), in which case the assignment process is straightforward. However, in instances where the codes of the personal name and the family name are different, it is necessary to apply a set of rules to establish which of the two names is the more reliable indicator of the origins of that individual. This is done first by establishing for each personal and family name a 'confidence score', indicating the relative extent to which its name is associated with the origin it is assigned to. On this basis, while both 'Ernst' and 'Arnold' would be assigned the code 'German', the name 'Ernst' would have a much stronger association with Germany than would 'Arnold' just as would the surname 'Schwarzenegger' than the surname 'Beck'. Thus, in instances where the two parts of a consumer's name are associated with different origins codes, the consumer would typically be assigned to the code of the name with the higher confidence level.

Different rules need to be applied to international names

Different rules need to be applied to international names. For example, it makes sense if one is coding a Spanish name file to consider the name 'Antonio' to be Spanish, while when coding an Italian file it may be more appropriate to consider it as Italian. If an 'Antonio' is found on a Swedish file, then it would be appropriate for the name to be allocated to either Italy or Spain depending on the relative frequency of the name 'Antonio' in the two countries.

Names, ethnicity, culture and language

Subjectivity

Potential users of a name-based segmentation systems may naturally question how accurate such a system might be. Such a question itself begs the question of what should be the yardstick against which accuracy is to be measured. The segmentations used to describe foreign populations are themselves subject to a certain degree of subjectivity. When 'immigrants' first arrived, it was typical to categorise them according to their racial origins, if only because it was the different physical appearance of these newcomers that was most striking. Race was often conflated with country of birth or nationality. By contrast, today's public opinion focuses to a much greater extent on religion. Whether a person is Muslim is seen to be more relevant than whether that person originates from Pakistan or Turkey in determining the likelihood with which they will adopt peculiarly British characteristics. These differences are now more visibly recognised by behavioural features, such as the wearing of a veil or a turban, than by people's physical appearance.

The determination of the categorisation is made more complicated by the extent to which certain immigrant groups have experienced more than one migration. For example, are Ugandan Asians to be deemed African, on the basis of continent of birth, or Asian, on the basis of the continent of their original forebears? For the marketer, language may be as or more important a key to answering this question than either race or religion. Clearly, a categorisation based on the analysis of people's names will tend to incorporate elements of racial, linguistic and religious classification systems. However, it may be more appropriate to view the classification as a discrete and different form of classification, with its own specific advantages and disadvantages over the others, rather than a system that merely seeks to approximate to an illusory 'gold standard'.

Another challenge that a name-based classification can make to existing ways of viewing minority communities is that people vary in the degree to which they belong to any one category. The 2001 census revealed that persons of mixed race represented 1.31 per cent of all residents compared with 11.70 per cent of persons of a single non-British ethnic group. The percentage of persons of mixed race tends to be much higher among better-off minority populations, particularly those originating in part from other European countries.

Assimilation

Additionally, the process of assimilation leads many population groups to adopt the consumption habits of the host population. Thus, fifth-generation Irish immigrants living in Liverpool or Middlesbrough are likely to describe themselves as English on census night and the Poles who emigrated to the Nottinghamshire coalfield in 1945 are likely to have very different dietary habits from those who have arrived more recently via Easyjet and Stansted airport. However, not all minorities assimilate at the same speed.

The names parents give their children, it could well be argued, reflect not just their origins but the extent of their assimilation. Thus, although descendants of Black Caribbean immigrants may still look very different from the host population, the fact that their language and religion are the same as those of the host population results in lower levels of cultural differentiation than is the case with Black African immigrants. The much higher proportion of persons whose names contain both a Swedish part and an English part that contain both a Sikh part and an English part shows that there is a much greater degree of similarity between Swedish and English cultures than between Sikh and English ones.

Results

To validate the processes used in building the origins coding tool, Origins was fortunate to be allowed to access a copy of Experian's UK Consumer Dynamics Database. This contains a list of all persons who have not ticked the opt out box on the UK electoral register plus the names of company directors, many shareholders and other names on publicly accessible databases. Altogether some 46 million names are found on this database.

Table 4: Names and ethnicity/religion: A comparison

Basis of segmentation	Personal and family name	Ethnicity/religion	
Universe	Adults	Population	
Source	OriginsInfo	census	
Population segment	%	%	Index
African	0.31	0.32	95
East Asian	0.35	0.43	81
European	1.44	1.13	128
Greek Orthodox	0.23	0.19	118
Hispanic	0.38	0.40	95
Jewish and Armenian	0.17	0.47	37
Muslim	2.11	2.22	95
Nordic	0.11	0.09	117
Sikh	0.59	0.59	101
South Asian	1.00	0.98	102

'Onomastic' groups

Table 4 shows the degree to which names on this file are distributed by the major 'onomastic' (ie name based) groups and how this compares with the closest estimates one can obtain for corresponding categories from the 2001 census. The comparison shows that the proportion of names in each main group generated by this method is broadly similar to the proportions of the UK population in each group with the exception of the Jews and Armenians, many of whom changed their names after arriving in Britain. If the table were to break down the categories in more detail, it would show there were significant discrepancies in terms of representation of the Japanese, few of whom may register to vote, and of Black Caribbeans, most of whom have always had British names.[15]

Reliability of assignments varies according to a person's origins

How reliable the assignments are likely to be varies according to a person's origins. As can be seen from Table 5, the proportion of people with an English personal name varies significantly between cultures. Groups such as Ethiopians and Basques as well as, more obviously, Bangladeshis and Hindu Indians are likely to be very effectively identified through such a system. In other words, each of these origins categories tend to have distinctive names. Black Africans, Chinese and Japanese also have distinctive names, though less so than the Sikhs, Hindus and Muslims. Among Europeans, people of Hispanic and Greek Orthodox origin are much more likely to be accurately coded than people of Scandinavian, German and French origin. The reason for these groups being less easy to distinguish is partly because of higher levels of inter-marriage, partly the greater similarity in the language and thirdly the greater tendency to adopt names from these countries when naming children.

That most culturally different minorities are the ones with the most distinctive names and the ones it is easiest correctly to classify is very convenient because these are the groups whose behaviours are most distinctive and who marketers, one would suppose, are most interested in reaching.

Table 5: Ranking of selected population groups by proportion with an English personal name

Origin of surname	% with an English personal name
English	91.3
Jewish	78.4
Black Caribbean	74.0
German	70.7
Hungarian	63.2
Basque	20.0
Turkish	16.5
Ethiopian	13.6
Sri Lankan	13.2
Indian (Hindu)	6.0
Bangladeshi (Muslim)	5.0

Profiling customer files

The process of gaining insight into the groups under- or over-represented on a client file is broadly similar in form to the process of coding and then profiling customers by type of neighbourhood. First, the client file is coded using the matching software and its rules for assigning names where the origins codes of the personal name differ from those of the family name. Then, the total number and proportion of names assigned to each type is compared with the number and proportion of names belonging to each type on a 'base' master file. Finally, the difference between the two distributions is expressed in the form of 'index' values.

Categories should really be considered as building blocks

One difference between the two forms of segmentation is that the 200 detailed categories into which names can be segmented is far too many for general use, particularly since many of them have very small frequencies. Thus, the categories should really be considered building blocks providing the flexibility to summarise results on a number of different dimensions. For example, for some uses it may be appropriate to organise the 200 categories on the basis of broad geographical region. For other applications it may be more appropriate to organise the 200 microsegments on the basis of religion and for others it may well be language that is the more important criterion. For example, a person with the name 'Mascarenhas' can be identified as belonging to Goan ancestry. This places them as South Asian in terms of their geographical origin, Portuguese in terms of language and culture and Catholic in terms of religion.

However, any use of names for profiling needs to address with care the manner in which the UK base population is defined. Clearly the names on the electoral roll or UK Consumer Dynamics Database are not totally representative of the country as a whole. Indeed it is often difficult, when measuring the size of particular market segments, to know which population groups one would ideally wish to include in a base against which one might want to compare one's customer records. Does one want to include temporary business visitors from Japan, as the census would, within one's count of Japanese people available to market to? Would one want to include full-time students, au pairs,

seasonal agricultural labourers, etc? What constitutes the appropriate 'base' for use in profiling will clearly vary from user to user and from application to application.

Clearly the problem of measuring the base population is less problematical when the customer profiling involves examination of behavioural differences within a file rather than when comparing the mix of segments on a file with the mix in the population at large.

Selections for targeting onomastic groups

Most segmentation systems are composed of categorical and mutually exclusive segments. At any one time a customer is either in a segment such as a 'Corporate Chieftain' or a 'New Urban Colonist'. Customer segments can normally either be selected or deselected.

Names-based classification offers increased variety of customer selection options

A names-based classification offers a greater variety of customer selection options. Consider the customer names listed in Table 6. These are actual names from the UK Consumer Dynamics Database that appear, on the basis of the algorithms used by the Origins coding system, to belong to people originating from Myanmar. If we want to be absolutely sure to target only those customers who originate from Mynamar, even at the expense of reducing the size of our selection, then we may wish to adopt the strategy of only targeting customers both of whose name elements come from Myanmar. This would clearly eliminate San Naidu and Tun Williams, both of whom have Myanmar personal names but appear to have married people from different cultures.

Alternatively, we may wish to expand our selection by including all names where either one or the other component of the name originates from that country, for example by targeting anyone with the personal names 'San' and 'Tun' irrespective of their family names. A further option is to apply a score-based cut-off, in other words to select customers who we believe are most likely to originate from Myanmar because their name elements are especially strongly associated with

Table 6: Some people with names originating from Myanmar (Burma)

People with low scoring names	Score	People with high scoring names	Score
SAN…NAIDU	0	SEIN…NGWE	8.51
TUN…WILLIAMS	0	AUNG…HLAING	8.52
JOSEPHINE…TOE	0.01	MAUNG…HLAING	8.53
LEONA…MOE	0.01	AYE…NGWE	8.60
MAUNG…SAW	0.01	LWIN…HLAING	8.78
WIN…GILL	0.01	AUNG…THANT	8.80
MARLIS…ZIN	0.02	KYAW…THANT	9.04
FREDERICK…WIN	0.03	KYI…HLAING	9.04
GEOFFREY…LATT	0.03	WIN…HTUT	10.31
KHIN…MURPHY	0.03	KHIN…HTUT	11.19
PAULA…THEIN	0.03	THAN…HTUT	11.20
SHAN…SHWE	0.03	AUNG…HTUT	11.28
SOE…JOWES	0.03	MAUNG…HTUT	11.29
ANGELA…THEIN	0.04	ZAW…HTUT	11.69

that country. In this case, we would perhaps select only the names in the right-hand list and deselect those in the left-hand list.

The option of selecting customers of apparently mixed origins may also be rewarding because these customers may be less rooted in their home culture and may have more reasons than other customers to respond to direct communications in the absence of the reinforcement of their cultural identity through interaction with other family members.

Aggregation of data to higher geographies

Statistics on the frequency distribution of adults by origins codes right down to postcode level can be useful

As was reported in the review of the building of Mosaic covered by *Interactive Marketing* in its January/March 2004 issue,[16] there is precedent for the use of names in market segmentation systems insofar as the proportion of Asian names by postcode was included in the list of data characteristics used in the most recent build of Mosaic. Given the level of residential segregation of minority groups in modern Britain, statistics on the frequency distribution of adults by origins codes right down to postcode level can be very useful, and there are some arguments for using this in preference to published census statistics for understanding local demographics. One of the reasons clearly is the finer level of granularity achieved by using the postcode rather than the census output area as a geographical unit. The second benefit is that the method avoids the randomisation which is imposed by the census where data for sparse groups are reported at a low level. Of equal importance is the fact that the information is updatable on an annual basis which, given both the increase in and movements of minority populations since 2001, may make a considerable difference to the accuracy of census-based neighbourhood statistics.

On the other hand, it is important to recognise when using this approach that the statistics will be much more reliable for settled, permanent populations who are more likely to be recorded on the electoral register than it will be for temporary residents, especially those from countries outside the European Union who are not eligible for inclusion on the electoral roll.

Targeting by context

Stereotypes do not work in practice

For reasons of ignorance rather than malice, it is easy to suppose that the minority populations one may wish to reach conform to a bland stereotype in which all members of the group are deemed to share identical attitudes, values and aspirations. Clearly in practice this is not the case. Just as there are wealthy Britons and poor Britons, so too the Muslim community is divided between extremely wealthy Arabs resident in Mayfair and extremely impoverished migrants from rural Bangladesh, many of whom live in Tower Hamlets. Indeed the income divisions between rich and poor could well be sharper among many of these minorities than among the host population. Thus, often it may be effective either to use names-based segmentation in combination with Mosaic or to use it in combination with the same data added up by postcode geography.

If, for example, we use name data to organise postcodes according to whether they are almost exclusively White British, whether they are mixed or whether the population of just one minority group predominates, we are then able to further segment a market group based on names so as to differentiate those members of the community who appear, residentially at least, to have integrated with the host population from those who live within a mono-cultural neighbourhood. These differences are likely to be especially important, not just in terms of product preferences but also in terms of access to and use of inter-personal face-to-face networks for the purchasing of products relevant to that ethnic group. Put simply, the opportunity to purchase a saree is much less if you live in a middle-class white suburb in Plymouth than if you live in the northern suburbs of Leicester.

Use of names-based segmentation together with geodemographics makes highly targeted communications possible

The use of names-based segmentation in combination with geodemographics makes it possible for the first time to undertake highly targeted communications with the most successful members of particular communities, many of whom are likely to have very distinct consumer needs.

The use of names linked to business registers

It used to be said of Victorian London that the Irish ran the building industry and also the pubs. The same is probably equally true today. Other groups have not only colonised their own residential neighbourhoods but also appropriated various business sectors (as well as manifestations of criminality), often for reasons that are not immediately apparent.

Names on registers can elucidate urban myths about business sectors

The presence of names on registers of company directors, partners and sole traders now makes it possible, probably for the first time, to identify the extent to which urban myths about the business sectors monopolised by the Irish, and now indeed their successors, are really true. Table 7, which is based on Experian's Business Information database, shows that post offices are disproportionately run by South Asians, that the manufacture of ice cream is still the preserve of Europeans and that people of Jewish origin are disproportionately likely to be involved in the sale and purchase of real estate. Indeed, the data suggest people of foreign origin are proportionately more likely, bearing in mind their share of the total population, to be directors, partners or sole traders than people of British origin.

While the precise value of this linkage has yet to be properly understood, one would suppose that the information should, for the first time, make direct communications a useful opportunity for targeting minority groups, whether for suppliers of business services, government incentive schemes or indeed high-end consumer products.

Ethical considerations

No discussion of the use of names as a basis for market segmentation could be properly concluded without some consideration of ethical standards and their policing.

Table 7: Origins of UK business owners and directors by industrial sector

Origin of name	Total directors, partners and sole traders	National Post Office	Manufacture of ice cream	Buying and selling of real estate
	%	%	%	%
English	59.7	55.7	46.1	56.4
Other European	3.5	3.7	22.7	3.5
Jewish	0.7	0.1	0	3.3
South Asian	5.2	17.3	7.3	9.2
Other	30.9	23.2	23.9	27.6
Total	100.0	100.0	100.0	100.0

General presumption within government that data on ethnicity should be captured wherever practical

The view of the author is that there can be no ethical objection, in principle, to the use of ethnicity as a basis for segmentation and targeting. Indeed there is a general presumption within government that, wherever practical, data on the ethnic origins of citizens should be captured by those public sector organisations responsible for the delivery of services to the citizen.

In some instances, as for instance with the Home Office and DfES, the information is specifically requested and captured in order to monitor the experience of ethnic minority users of services, as for instance where using the PLASC database DfES monitors the average performance of pupils from different minorities at each key stage or the Department of Heath monitors hospital diagnoses by ethnic group.

While at one level the information is used for research and monitoring purposes, government is increasingly moving towards an approach where units responsible for delivering particular services are required to demonstrate that they are reacting to differential risks among particular minority populations by targeting particular communications at them. Likewise they are increasingly enjoined to deliver services in such a way as to reflect the cultural sensitivities of users from key minority groups. This practice is particularly advanced in the public health campaigns, where it is considered appropriate to use ethnicity data as a basis for 'social marketing'[17] and for the targeting of messages as well as treatments.

Ethnicity data as a basis for 'social marketing'

Applying the principles used in 'social marketing' to the marketing of commercial products, it would seem quite appropriate therefore to target information on products specifically developed to meet the needs of particular minorities to the members of those minorities for whom they have been developed. Thus, it would seem wholly consistent with practice in the government sector to target information on Asian foods in a supermarket to loyalty card holders who were of Asian origin, to target information on Sharia compliant savings products to bank customers who were likely to be of Muslim origin and for loyalty schemes to offer as incentives to Greek Cypriot customers free flights to Paphos rather than to Miami.

Nevertheless, there are potential applications and processes which members of all communities would consider as inappropriate. These

would include applications that contribute to further residential segregation or that recruited individuals to organisations whose aims were to reinforce divisions between cultures or to incite hatred.

Likewise, it would be inappropriate to use names as a basis for determining the language in which a communication is written, although it may be appropriate as a basis for offering a wider variety of language and channel options, as for example in many statutory communications emanating from local government. It is for these reasons that Experian, the principal provider of the Origins segmentation system, incorporates a specific code of conduct covering these points in the contract which its clients are required to sign.

References and Notes

1. ONS neighbourhood statistics. http://www.statistics.gov.uk/census2001/census2001.asp.
2. *The Sunday Times*, Sunday, 10 September 2006, p. 13.
3. According to the Evening Standard, 27th October, the 29 new appointees bear 20 non-British surnames: Boomars, de Pourtales, Ekaireb, Ferrari, Flynn, Grabau, Irani, Malhi, Oppenheimer, Pantazopoulos, Rao, Saidenburg, Sotir, Stranger, Sze, van Praag, von Kuskull, Weber, Wisnia, Zaimi. Only nine, Beveridge, Burgin, Faker, Holder, Mansfield, Metherell, Selman, Wilson, Wright, were of British origin.
4. Mosaic Multimedia Guide, available from Experian, Talbot House, Talbot Street, Nottingham, NG80 1TH.
5. http://www.originsinfo.com.
6. http://www.originsinfo.com.
7. 'How to Market Better Health — A Dr Foster community health workbook', *Dr Foster Intelligence*, November 2004.
8. Sleight, P. (2004). *Targetting Consumers, Second Edition — How to Use Geodemographic and Lifestyle Data in Your Business*, WARC, Henley on Thames.
9. *The Daily Telegraph and The Times*, Monday, 16 October 2006.
10. Cummins, C., Winter, H., Cheng, K., Maric, R., Sicocks, P. and Varghese, C. (1999) 'An assessment of the Nam Pehchan computer programme for the identification of names of South Asian ethnic origin', *Journal of Public Health Medicine*, Vol. 21, No. 4, pp. 401–406.
11. Nanchahal, K., Mangtani, P., Alston, M. and Dost Santos Silva, I. (2001) 'Development and validation of a computerised South Asian name group recognition algorithm', *Journal of Public Health Medicine*, Vol. 23, No. 4, pp. 278–285.
12. Hanks, P., (ed) (2003). *Dictionary of American Family Names*, Oxford University Press, Oxford.
13. Tucker, D. K. (2005) 'The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN)', *Onomastica Canadiana*, Vol. 87, No. 2, 71–84.
14. Tucker, D. K. (2005). *The Changing Faces of the UK*, ICOS XXII Conference, Pisa.
15. Interestingly, a disproportionate number of Black Caribbeans bear Scottish and Welsh family names, a reflection of the origins of the plantation owners from whom they typically took their names.
16. Webber, R. (2004) 'Designing geodemographic classifications to meet contemporary business needs', *Interactive Marketing*, Vol. 5, No. 3, pp. 219–237.
17. 'Reaching People — Social Marketing in Practice', *Dr Foster Intelligence*, September 2006.