Exploring 'invisibility' in China's Digital Economy

Ashley D. Lloyd
Business School
The University of Edinburgh
29 Buccleuch Place, EH8 9JS
+44 (0) 7914 895108
ashley@edinburgh.ac.uk

Mario A. Antonioletti
Terence M. Sloan
EPCC
The University of Edinburgh
James Clerk Maxwell Bldg, EH9 3JZ
+44 (0) 131 650 5141/5155
m.antonioletti@epcc.ed.ac.uk
t.sloan@epcc.ed.ac.uk

Yvonne Barnard
Institute for Transport Studies
University of Leeds
36-40 University Road, LS2 9JT
Telephone number, incl. country code
Y.Barnard@leeds.ac.uk

ABSTRACT

Access to products and services is increasingly 'digital by default'. Non-users of digital channels have no direct means of signalling preferences through consumption, rendering them effectively invisible to designers of future products and at risk of permanent exclusion. We explore this in China by sampling millions of consumer's records across digital and non-digital channels. Separating consumers into digital/non-digital user groups allows the characteristics of those at risk of 'digital exclusion' to be predicted. From a corporate perspective, this scale of analysis allows strategic models of economic development in China based on both City Tier and McKinsey's City Cluster to be tested, and delivers 'at risk' groups that are large enough to provide economic incentives for inclusion. Non-use may, however, be elective, with consequences for the effectiveness with which governments formulate and target 'digital inclusion' policies. Hence we also explore how elective non-users might be distinguished from the potentially excluded.

Categories and Subject Descriptors

J.4 SOCIAL AND BEHAVIORAL SCIENCES; J.1 ADMINISTRATIVE DATA PROCESSING: Business, Financial, Marketing; H.2.8 Database Applications: Data mining

General Terms

Management, Design, Economics, Human Factors, Theory

Keywords

China, Big Data, e-Social Science, Digital Exclusion.

1. INTRODUCTION

In a global digital economy, the relative ease with which producers can deliver through 'direct channels' means that the physical distances that separate producer and consumer can be of the same scale as the market, i.e. global [1]. As direct producer-consumer distance increases there is also a tendency to reduce the range of customer interactions, leaving producers increasingly reliant on 'Big Data' about their customers collected through the same channel through which products and services are consumed. Parts of society whose needs are not reflected in the design and delivery of current products cannot have their preferences tracked by monitoring consumption and become effectively 'invisible' to the designers of future products. This is not good for society, but nor is it good for industry seeking to grow new markets, where inclusion is served by understanding user needs better, innovating through good design and delivering well-targeted products and services on a scale that minimises economic barriers.

In this paper we report an RCUK-funded project addressing both visibility and scale issues by establishing an analytical facility within a Chinese data centre dealing with more consumers than the UK population across both digital and non-digital channels.

2. METHOD

2.1 Definitions: Invisibility vs. Exclusion

Non-users are 'invisible' to a digital economy where consumption establishes the principal means of communication. Non-users are not necessarily excluded however, as they may be electing not to use these technologies. This distinction allows us to consider how these types of non-use might be distinguished analytically.

We must also take care to define how use leads to a 'visible' engagement with the digital economy. Visibility is about identifiable use and hence we define 'visible' use as that which explicitly establishes a personally identifiable digital trace in the consumption record. For example, a consumer may use Mail or the Internet to order the same product using the same payment method. In both cases a data trace is established in the order processing system, but it is only by using the Internet that a consumer synchronously establishes and validates that trace. With Mail order the company validates the record and hence a Mail order user is classed as a member of the 'non-digital' group.

2.2 Data Access

This project builds on the ESRC eSocial Science demonstrator INWA that first established a UK-China Grid in 2004 with the Computer Network and Information Center of the Chinese Academy of Sciences [2][3]. Following extensive co-design to address security and management requirements, and testing for ISO 9001 compliance, we installed a facility within a Chinese corporate data centre to gain access to the consumption data of millions of consumers over a period of years, across channels that ranged from shops and mail order to online sales. This breadth is also reflected in the products and services being consumed, from branded highly differentiated products to generic consumables, giving access to a wide range of consumer demographics.

2.3 Data Mining

After applying statistical techniques to clean the data we interviewed business analysts and senior management to test our

data abstraction of the organisation's business processes and identify expected behaviours. We also interviewed staff in a shop setting to determine the consumer behaviour they observed.

This enabled us to separate the population into behavioural types and apply a data-mining tool, C5.0 [4] to discover whether these types of behaviour were associated with any patterns within the data that would allow classification of users and non-users of digital channels in terms of distinctive socio-demographics [5][6].

C5.0 is a highly scalable development of C4.5, ranked as the most influential data-mining algorithm at the IEEE International Conference on Data Mining in 2006 [7]. C5.0 splits the sample according to the attribute that provides the most information gain, then splits the subsample by the attribute that provides the most information gain and repeats until the sample cannot be split any further. Candidate trees are then pruned using heuristics derived from the Minimum Description Length Principle, which may be thought of as a cautious application of 'Occam's Razor' [8]. This results in smaller trees that are not only expected to have greater predictive accuracy, their structure is easier to follow and the segments (the population represented by each 'leaf') showing similar behaviour are necessarily larger. From an economic perspective such segments represent distinctive markets for which product/service design criteria are different, and the more each segment can be grown, the more viable it is to address.

2.4 Data Models: McKinsey City Cluster

The C5.0 models are developed to 'explain' the observed behaviour, however to make them of practical value in a corporate context it is important to establish links with other approaches to decision-making about expansion via economies of scale or scope [1][9]. Established models of economic development in China include the City Tier: a 'league table' of Chinese cities that can be extended to over 800 Chinese cities, from 'Tier 1' cities such as Beijing, to Tier 3 cities such as Bayan Nur. This has been criticised for its historical weighting and lack of reflection of growth rate, leading to an alternative proposition from McKinsey and Company: the City Cluster [10][11]. McKinsey divide China into twenty-two city clusters ranked by size: Mega, Large, Small and None. This approach is claimed to better reflect current consumer behaviour and hence adding the Cluster in which a consumer is located would be expected to explain more of the variation in behaviour in the final C5.0 model than City Tier.

3. DIGITAL CHANNEL USE IN CHINA

In the present study of Channel adoption in China, the decision tree resulting from the analysis was found to correctly classify 74.4% of 'Digital' users and 77.8% of 'non-Digital' users. This model of Channel adoption is abstracted in Figure 1, showing the most significant behavioural splits at the top of the decision tree, where we see that:

(I) The attribute that explains splits in the behaviour of the population better than any other is shown in the root node: the 'Cluster' in which the individual is economically active.

(II) The next most significant attributes focus on consumption volume and value, with higher order value/number explaining preferences for Digital channels rather than Non-Digital channels.

(III) It is only relatively low in the abstracted decision tree where cities are not part of identified clusters, that we see classical demographics of Education, Gender and Age explaining behavioural splits in the population.

This pattern is reflected in the full decision tree, containing 398 rules, within which overall attribute use is shown in Table 1.

4. DISCUSSION

The summary of attribute usage in **Table 1** details the relative significance of each attribute in explaining the observed behaviour. It is reasonable to infer from this table that the McKinsey Cluster does provide more information value in this case than the City Tier – i.e. that similarity in observed behaviour is better characterised by the Cluster in which a city is grouped than it is by the Tier in which the city ranked. Whilst this supports McKinsey & Co.'s [10] contention that City Cluster has replaced City Tier as the schema that explains most economic behaviour, they also note from longitudinal analysis that not all aspects of behaviour follow a Cluster model better than a Tier model. This explains City Tier's inclusion lower down the table, but what is notable is that City Tier appears higher up, and hence contains more information value, than traditional demographic attributes.

A question raised by this dominance of geography in the model, is whether the lower levels are picking up behavioural characteristics that will dominate as Chinese markets mature, behaviours converge, and the influence of geography on consumption reduces. McKinsey [11] point to the possibility of looking for such trends in the population as a whole through cross-sectional comparisons of clusters that show marked differences explained by their relative maturity: *"developed-country behavior patterns tend to be much more prevalent in clusters along the coast - where the economy has been prospering for a much longer period of time"*.

To explore this we can extract from the overall 398-rule decision tree, the branch relating to a Tier 1 coastal city that is part of a 'Mega' Cluster, and compare it to the 'Not in Cluster' ruleset as this samples predominantly rural cities (**Figure 2**). Such comparisons should be approached with caution as the samples reflected in the lower levels of a model necessarily represent smaller, and potentially less representative, groups. However in both rulesets we see patterns of non-use that might be explained by factors traditionally associated with exclusion: 'X-Ed' - Education/Experience and 'X-Ec' - Economic. Of particular note however are non-user groups with a counter-intuitive dependency on traditional demographics such as 'X-L1?' and 'X-L2?' where the model predicts Non-Use of digital channels when Education and the Volume or Value of transactions are in the highest category, reversing the pattern established for the majority of the sample in (II) above.

On this basis we might suggest that this non-use of a Digital Channel in a part of our sample that is highly educated and amongst the economically most active is *elective*. In other words, that the segments of the population represented by 'X-L' are by definition 'Digital Invisible', but not 'Digital Excluded'. They represent a potentially significant market segment for a supplier, but will not be 'made visible', i.e. adopt these technologies, simply by the returns to scale that reduce economic barriers. Nor are they likely to change their preferences as a result of government policy interventions intended to address access to infrastructure or other barriers such as education, as neither of these represent barriers for these 'X-L' segments.

Though this work is exploratory we conclude that an effective 'digital by default' agenda needs to understand the X-Ec/X-Ed/X-L boundaries to distinguish between 'exclusion' that can be solved by policy interventions versus that which might be solved by market processes, made viable by scaling up to global markets.

EXHIBITS

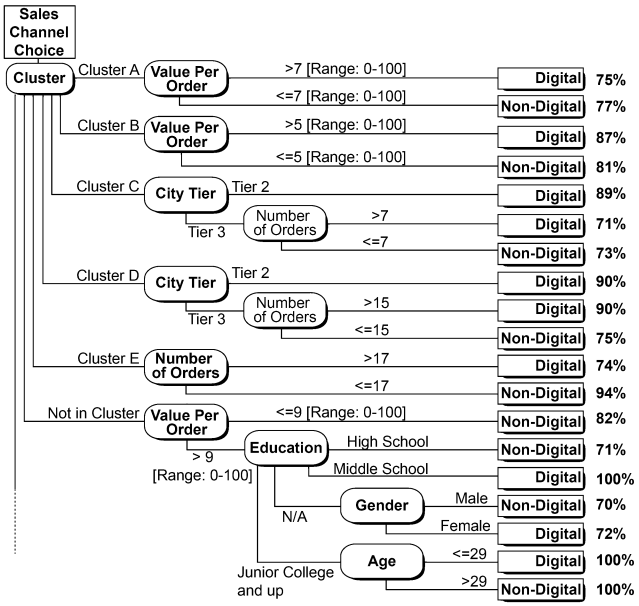

Figure 1: C5.0 Decision Tree Model of Digital/Non-Digital channel use. Graphic shows the top of tree showing major population splits and the relative weight given to City Cluster over City Tier and consumer demographics.

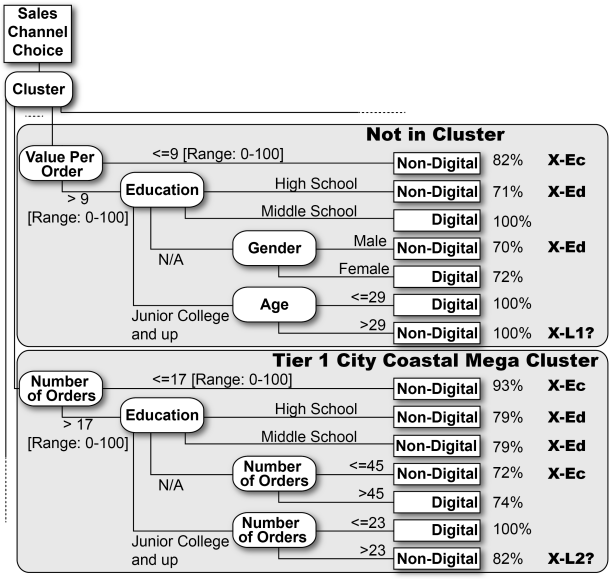

Figure 2: Abstraction of Decision Tree contrasting 'Not in Cluster' branch from Figure 1 and 'Tier 1 City within Coastal Mega Cluster'. Each leaf node for Non-Digital Users is tagged with a code indicating a reason for non-use suggested by the decision path: X-Ec = exclusion on Economic grounds, X-Ed = exclusion on Educational (Experience) grounds, X-L1?/X-L2? = possible elective non-use.

Table 1: C5.0 Model of Digital Channel Use: Attribute Usage

Attribute usage:			
100%	Cluster	21%	Occupation
89%	Number of Orders	17%	Age
64%	Value Per Order	10%	Education
42%	City Tier	8%	Gender

5. ACKNOWLEDGMENTS

The authors are grateful for the support they have individually and collectively received from a number of initiatives, from which we highlight a decade of support from the UK Economic and Social Research Council (award RES-149-25-0005, two tranches) for the initial phases of the INWA Grid and 'Follow-On Funding' (award RES-189-25-0039) and the EPSRC under the Digital Economy Programme (award EP/H006753/10).

6. REFERENCES

[1] Teece, D. J. 1993. The Dynamics of Industrial Capitalism: Perspectives on Alfred Chandler's Scale and Scope. *J. Econ. Lit.*, 31, 1, 199-225.

[2] Lloyd, A.D. and Sloan, T.M. 2011. Intercontinental Grids: An Infrastructure for Demand-Driven Innovation. *J. Grid Comput.*, 9, 2, 185-200.

[3] Lloyd, A.D., Sloan, T.M., Antonioletti, M.A. and McGilvary, G. 2013. Embedded systems for global e-Social Science: Moving computation rather than data, *Future Gener. Comp. Sy.*, 29, 5, 1120–1129.

[4] Rulequest Research C5.0. 2012, Available http://www.rulequest.com (Accessed 28 June 2013).

[5] Hume, A.C., Lloyd, A.D., Sloan, T.M. and Carter, A.C. 2004. Applying Grid Technologies to Distributed Data Mining. In *Proceedings Grid and Cooperative Computing 2004*, LNCS Vol 3251/2004, Springer-Verlag, Heidelberg, 696-703.

[6] Lloyd, A.D. 2005. The Grid and CRM: from 'If' to 'When'. *Telecommun. Policy*, 29, 153-172.

[7] Wu, X., Vipin K., Quinlan J.R., Ghosh J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z-H, Steinbach, M., Hand, D.J., Steinberg, D. 2008. Top 10 algorithms in data mining, *Knowl. Inf. Syst*., 14, 1–37.

[8] Quinlan, J.R. and Rivest, R.L. 1989. Inferring Decision Trees Using the Minimum Description Length Principle. *Inform. Comput.* 80, 227-248.

[9] Chavas, J-P. and Kim, K. 2010. Economies of diversification: A generalization and decomposition of economies of scope. *Int. J. Prod. Econ.*, 126, 229–235.

[10] Atsmon, Y., Ding, J., Dixit, V., Leibowitz, G., Magni, M. and Zipser, D. 2009. *One Country, Many Markets – targeting the Chinese Consumer with McKinsey ClusterMap*. McKinsey Asia Consumer and Retail. McKinsey & Co., Boston.

[11] Atsmon, Y., Magni, M. and Lihua, L. 2012. *From Mass to Mainstream: Keeping Pace With China's Rapidly Changing Consumers. Annual Chinese Consumer Report.* McKinsey Consumer & Shopper Insights. McKinsey & Co., Boston.