FRONTIERS OF MARKETING DATA SCIENCE JOURNAL

At the Forefront of the Marketing Data Revolution

THE INAUGURAL I-COM DATA SCIENCE JOURNAL

Methodologies, Applications, Models and Analysis

ISSUE No.1 | 2018

i-com
Global Forum for Marketing Data and Measurement

H2C
HO
HO
OH
O
R
R

CATALINA® buyR³science™ Relevant. Real Time. Results.

A Look At Catalina by:
Wes Chaar, Chief Data and Analytics Officer
Marta Cyhan, Chief Marketing Officer

Catalina is a big data company sitting at the intersection of marketing and technology. Having invented one-to-one marketing in 1983, literally no other company has amassed as much consumer data. This gives us a uniquely rich and expansive view of consumer buying behavior. We've branded our approach "buyR³science: Relevant. Real Time. Results." Both on-line and off. We know how to reach the ideal audience, communicate relevant value, engage shoppers at the optimal moment, reduce waste, and segment by price and deal sensitivity. As an organization, we deeply believe there is a science behind every buy, and a unique buyer behind the data.

Our Data & Analytics team plays a critical role here by driving a Machine Learning and AI transformation at the company, monetizing the inherent wealth of our massive shopper database. Meanwhile, our talented team of Data Scientists delivers always-on predictive and prescriptive analytics models to support critical business processes such as revenue management, multi-touch attribution, churn modeling, advanced personalization, and more. Using advanced Machine Learning and AI, we're processing Big Data on the Cloud at scale.

Our models help marketers understand how digital impressions relate to store purchases, where advertising dollars can best be spent, and what the drivers of consumer response are. We project what campaigns can be expected to deliver, and automatically close the gap when there are discrepancies – making Catalina a valued and valuable partner to the thousands of major and emerging brands, retailers, and agencies we serve, while providing the end consumer with attractive personalized offers.

By delivering our customers a true 360-degree understanding of shopper needs and preferences, we are able to reach the right households with highly relevant content in real-time, whether in-store, online or via their mobile devices. Nothing less than truly Optimized Marketing is our end goal – creating a world where our customers sell more and spend less.

Are you a Data Scientist? Come join our team and apply your analytics talents here. You won't find a better or richer digital and store data playground than what we offer at Catalina.

Table of Contents

The I-COM Primer on Blockchain's Application to Advertising Technology

Joshua Koran[1]
Sizmek

Richard Bush[2]
NYIAX

Jean-Paul Edwards[3]
OMD

Luke Mulks
Brave Software

[1] *Member of: I-COM Data Science Hackathons Board,
Data Creativity Awards Board and Jury,
Data Startup Challenge Board and Jury,
Journal Editorial Board,
San Francisco Advisory Board,
Co-Chair of: Data Privacy Council,
Blockchain and Advanced Research Council*

[2] *Member of I-COM Artificial Intelligence Council*

[3] *Member of I-COM Blockchain and Advanced Research Council*

Classifications, Key Words:

- Blockchain
- Discrepancy Management
- Workflow Automation
- Fraud Detection
- Verification
- Identity Management

Abstract

Blockchain is increasingly being mentioned in media, yet there is still a wide gap in understanding what it is and how its applications can benefit marketers. This paper provides an overview of the key technology components and some of the challenges and benefits of applying blockchain technology to marketing, both at the ecosystem-level as well as at the real-time transaction-level. While blockchain holds tremendous potential to improve negotiations, transaction auditing and supply chain management, given the current processing cost associated with the high volume of advertising transactions, widespread adoption is still a way away.

1. Technology Overview

Blockchain is increasingly being mentioned in media, yet there is still a wide gap in understanding what it is and how its applications can benefit marketers. This paper provides an overview of the key technology components and some of the challenges and benefits of applying blockchain technology to marketing, both at the ecosystem-level as well as at the real-time transaction-level.

Blockchain refers to a set of distributed technologies that ensure immutable storage of information, despite the lack of trust among the actors who write to this common ledger. The information is stored in "blocks," which are linked to each other in a "chain." The immutable property is provided by hashing information from the current block with information from the previous block. Because the network of participants validates each new block, as the number of participants grow, the harder it is for any group of participants to change previously recorded information. This is the chief reason behind blockchain's so called immutable property.

The validation of new blocks relies either on a proof-of-work or proof-of-stake mechanism. Proof-of-work requires participants to solve complex mathematical problems, creating an indirect transaction cost that limits the ability of a small group of participants to control the ledger. Proof-of-stake weights the votes of participants by their underlying ownership in the blocks being written.

Any alteration in the underlying blocks create a fork in the chain. Hence, this common ledger provides a single source of truth of both - the current block and all previous blocks in the chain.

There are four main types of blockchains, based on whether they are publicly or privately viewable and whether all participants can validate blocks or whether the validation is restricted to a permissioned set of validators.

Another entity associated with blockchains are "smart contracts," which are pre-arranged rules that trigger new blocks based on specific events. One benefit of these smart contracts is their ability to automatically execute, without needing a separate authority to execute the contract. The smart contracts also have the ability to govern downstream dependent processes and relationships amongst multiple parties. Because smart contracts are part of the blockchain, the blockchain serves as an immutable ledger that both records each contractual step and automatically triggers the next agreed upon action or actions.

Since the blockchain is ever increasing in size, as new blocks are added to it, another variation of this technology is to store some of the information outside of the blockchain. For example, the blockchain could merely store a link to some external document. While this dramatically reduces the storage-size required by the blockchain, the immutable guarantee of the blockchain is limited strictly to this reference, rather than the content of the information referenced by the chain.

2. Application to advertising

In a distributed supply chain and fragmented marketplace, standardising on an agreed upon method of determining units of exchange, authenticating and accrediting exchange partners and determining what triggers the completion of a transaction benefits all participants. Whilst traditional markets often solve this through centralised authorities, blockchain offers a novel, decentralised approach to addressing each of these fundamental market needs.

Blockchain technologies are being applied to marketing at both - the workflow and transaction event levels. At the workflow level, blockchain technology can help buyers and sellers negotiate on the inventory or audience information being transacted. In this regard, blockchain can help with supply chain management.

The current supply chain for purchasing digital assets often involves multiple parties that can lead to a lack of transparency and potential for fraud. Blockchain based approaches are enabling decentralised networks of trust in areas such as whitelisting of sites and inventory that are voted on by other members of the network. Incentives are realigned as providers of high quality inventory seek to come together maximising their scale whilst maintaining levels of quality.

At the event level, this same metadata helps standardise what media is being put up for sale, the identities of the parties involved in the transaction and validating that the delivered media matches what was put up for sale. Whilst multiple systems use their own unique taxonomies to classify information; an additional benefit of the standardisation of metadata helps distributed parties use a common language to reduce discrepancies in reporting. Blockchain technologies are also being investigated as a means to authenticate various elements of the media transaction such as validating a view (connected to a validated consumer identity) and minimising fraudulent activities such as domain spoofing. The complexity of modern targeting and decisioning can be embedded within a blockchain so that advertisers can be sure that all of their investment is going to the right audiences with the most appropriate message.

An additional application of blockchain technology is to enable consumer identity management. To date, consumers are often asked to declare personal data in return for free services. Blockchain allows for stable, secure identity profiles to be controlled by the consumer via a smart contract with access given to trusted partners who can leverage that data in different ways. This may enable consumers to see fewer, and more relevant ads in exchange for the same free content. An example is a blockchain-enabled insight platform that allows different parties to come together to aggregate consumer data without being given in a raw format to another party. Taking identity management to the realm of the internet of things (IoT), blockchain technologies can manage the identity of devices.

For example, a dishwasher ID may be connected to a credit card ID to find the best deals in dishwasher tablets. Moreover, the identity of market participants can be validated by using standard public/private key encryption. The use of encryption can also enable market participants to control access to understanding the contents of given blocks within the public ledger. Of course, the applications to the advertising ecosystem are not limited to these uses. Whilst advertising participants can already communicate and transact without blockchain technologies, the primary benefit of blockchain is to provide trust among parties, without either party having to disclose their identity or rely on a centralised authority to validate that their transaction is complete.

3. Technical Challenges

One of the common challenges discussed in association with blockchain is the cost of processing each new block, as well as the latency involved in validating transactions. This processing cost is sometimes referred to as "gas." Most blockchain technologies can only process thousands of transactions per second. However, programmatic advertising requires technologies that scale to millions of transactions per second. This is one the reasons that many marketing-focused blockchain technologies are focused on addressing the workflow-level rather than transaction-level challenges. The initial blockchains that are well know from the cryptocurrency world have not been designed for the marketing world and therefore don't address the typical latency needs. One approach to solve this can be to secure the data written by aggregating and hashing it into a side chain or sharded architecture.

4. Ecosystem Challenges

One of the benefits of a common distributed ledger would be the ability to reduce discrepancies in counting. Given the different methodologies in counting users and events, the counts in different systems may differ by many percentage points. By agreeing to use a common system of counting, the various systems could hone their own internal mechanisms to all rely on this common system.

Another benefit of a publicly viewable distributed ledger is the promise that it can reduce fraud. By improving the transparency around the identity of sellers and buyers, as well as distinguishing robots from consumers, it would greatly improve the ability of fraudulent participants to take advantage of the current marketplace. To address the issue of fraud, this improved transparency would need to identify what is being transacted, such as the user and context of the media placements. One of the requirements to make this a reality would be for sellers to agree on a common taxonomy of inventory context and audience information, such that buyers could more easily validate what was bought against what was sold. Indeed, many buyers are asking for increased transparency into how their money is spent on inventory, data, technology and services.

However, not all market participants desire full transparency. Sellers face increased sales channel conflict with increased transparency. Moreover, many sellers are concerned about the leakage of their data such as which audiences frequent their websites or the price at which they sell their inventory. Similarly, most buyers do not want to share the marketing plans or prices they have negotiated with their competitors. Even value-adding intermediaries are reluctant to share how they use proprietary data to improve their match of content to consumers. Given the desire of many entities to not publicly disclose their custom data, their purchasing tactics and the amounts transacted, it is unlikely that this information will be made available in a publicly viewable blockchain.

The trading secrecy that both publishers and buyers have means that the information stored in the blockchain may itself be encrypted, such that only a limited set of participants will be able to read the information that is stored in the publicly viewable ledger.

Since user IDs associated with consumer applications or browsers need to be available to validate the delivery of media contracts, this brings up the issue of consumer privacy. Many consumers would prefer not to have all their browsing and application usage information made publicly available. Thus, there may be both self-regulatory and legal requirements to obfuscate the user ID associated with the transaction from the identity of consumers.

Conclusion

Blockchain technologies will continue to offer marketers a new ability to buy media and measure its effectiveness. In the short run, blockchain is most likely to help with workflow improvements. The performance improvements in processing technology will eventually enable blockchain to be used at scale at the transaction level.

This year, the I-COM Data Science Council on blockchain focused on a high-level overview of distributed ledger technologies to advertising ecosystem. We thank the entire Council for their contributions during our monthly meetings that made this article possible.

References

1. Various, "Proof of Stake," Accessed from https://github.com/ethereum/wiki/wiki/Proof-of-Stake-FAQ

2. Various, "A Next-Generation Smart Contract and Decentralized Application Platform," (2018) Accessed from https://github.com/ethereum/wiki/wiki/White-Paper

3. Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," (2008). Accessed from https://bitcoin.org/bitcoin.pdf

4. Bill Wise, Boris Mouzykantskii, "For Blockchain To Work In Ad Tech, We Need To Be Honest About Its Challenges," AdExchanger (August 29, 2017). Accessed from https://adexchanger.com/data-driven-thinking/blockchain-work-ad-tech-need-honest-challenges/

5. Eric Berry, "Is Blockchain The Best Solution For Ad Tech's Most Pressing Problems?" AdExchnger (August 22, 2017). Accessed from https://adexchanger.com/data-driven-thinking/blockchain-best-solution-ad-techs-pressing-problems/

6. Manny Puentes, "What Blockchain Can (And Can't) Solve For Ad Tech," AdExchanger (April 26, 2017). Accessed from https://adexchanger.com/data-driven-thinking/blockchain-can-cant-solve-ad-tech/

7. Various, "Blockchain for Video Advertising: A Market Snapshot of Publisher and Buyer Use Cases," IAB (February 2018). Accessed from https://www.iab.com/wp-content/uploads/2018/02/Blockchain_for_Video_Advertising_Publisher-Buyer_Use_Cases_2018-02.pdf

Authors

Joshua Koran is currently Sizmek's Managing Director for their data enablement product line. Joshua is a true pioneer in digital marketing, behavioral targeting, data marketplaces and data management platforms with nearly 20 years' experience in these domains.

AdWeek recently recognised his contributions to this field, in which he holds 22 patents, by naming him one of the top innovators in the space. Among his many accomplishments, he designed the industry's first predictive behavioral targeting platform, the first data marketplace ad network, and the first three-screen (TV, mobile, display) behavioral profiling platform. Over the course of his international career, he has lead product, engineering, data science, user experience, business development, and commercialisation teams in companies such as Turn, AT&T, Yahoo!, and ValueClick. Through his digital consultancy he has helped nearly two dozen other digital advertising companies improve their understanding and monetisation of online data. Joshua holds B.A. from Boston College, J.D. Law from U.C. Hastings College of the Law and M.B.A. from Oxford University.

As President of NYIAX, Richard Bush spearheads the ongoing development of the NYIAX platform. Richard has more than 15 years experience in the advertising technology industry. He comes to NYIAX from IPONWEB, a key infrastructure provider in the media trading ecosystem, where he was the General Manager of its Publishing Solutions business. In this role, he guided custom projects for industry players to market, and working with teams across the globe (Japan, EU and the USA) built new products and custom platforms specifically tailored to client's unique business models. Richard has also served as the VP of Product and Technology at AOL Networks, and as the original programmer and developer of content management and web publishing software for many of the Reed Business Information's web properties. Richard currently resides in the Greater New York City area.

Jean-Paul Edwards has been with OMD for 20 years. He founded Manning Gottlieb OMD's Digital team in 1997 and then led the agency Media Futures' offering. He now works at OMD EMEA to drive development of the network's offering in a digitally led, data-centric media environment.

Luke Mulks is the Director of Business Development at Brave Software, and is a core member of the Basic Attention Token (BAT) team. Prior to joining Brave, Luke was the Director of Ad Products at OAO, where he handled ad product integrations and operations for OAO's clients, who are among some of the largest media companies and publishers on the web. Luke's focus at Brave is to apply his expertise from his background in publishing, startups and advertising into the Brave browser and Basic Attention Token platform.

CONTRIBUTING COUNCIL MEMBERS

Alexander Czernay, Namics
Amit Phansalkar, Netra Systems
Andy Fisher, Merkle
Antonio Carbajal, Turner
Brian Fitzpatrick, IPONWEB
Dirk Beyer, Neustar
Eduardo Salazar, MacroAnalytica
Jake Sroczynski, Amnet Group US – The Dentsu Aegis Programmatic Experts
James Dutton, TrafficGuard
Jerome Shimizu, DoGood Media
Johan de Groot, Faktor B.V.
Jonathan Sweeney, Gain Theory
Julian Berman, Magnetic
Kai Xiang Neo, Amnet Programmatic Experts for Dentsu Aegis Network
Kajal Mukhopadhyay, Verizon
Ken Brook, MetaX
Krassimir Kostov, Cognizant Analytics
Lauren S. Moores, Indigo
Matti Parssinen, Cygate
Miguel Morales, Lucidity
Nukhet Kayahan, Deloitte
Peter Picado-Curtis, J.P. Morgan
Peter van Leeuwen, Fulcrum.Works
Praneet Sharma, Method Media Intelligence
Sandeep Jeereddy, BMW of North America
Sanjeevan Bala, Channel4
Shailin Dhar, Method Media Intelligence
Simon Honc, [m]PLATFORM at GroupM
Stacy Huggins, MadHive
Tim Geenen, Faktor B.V.
Walid Hadid, Decenture

I-COM Data Science Council on Blockchain and Advanced Research

The purpose of the Council is to share information on what blockchain is and its application in the Marketing and Advertising Industry, and to educate and provide constructive advice to leaders and practitioners of the ecosystem.

This initiative within the framework of I-COM Data Science Programme was launched in August 2017.

The main objectives of the group are to:

- Examine different blockchain technologies and how applications of that technology can improve our day-to-day business;
- Analyse how we can use contracts and maintain transparency and integrity;
- Explore how blockchain could solve different business problems in the industry, comparing it with other possible solution projects.

FRONTIERS OF MARKETING
DATA SCIENCE JOURNAL

Mixed Effects Marketing Mix Modelling Can Reveal Significant Heterogeneities in Advertising Response

Saeed R. Bagheri
Amazon Advertising

Seyed Hanif Mahboobi
Amazon Web Services

Mericcan Usta
Apple

Jing Zhao
GroupM

Hamid R. Darabi
Remedy Partners

Classifications, Key Words:

- Marketing mix modelling
- Marketing performance measurement
- Mixed-effect linear regression
- Optimum marketing budget allocation
- Econometric modelling

Abstract

How have sales responded to prior levels of advertising expenditures? A Marketing Mix Model (MMM) provides an econometric approach to generalise key performance indicators of marketing efforts for many large advertisers as well as many agencies and vendors that serve the advertising analytics needs of these advertisers. MMMs can cover fundamental response effects of advertising including carryovers, lags, and saturation. In reality, a typical campaign exposes diverse populations across multiple markets. Failure to recognise the heterogeneity in responsiveness to advertising may lead to misleading insights using such models. Mixed effects MMMs allow modelling for variations in responsiveness along multiple dimensions such as geographies. Mixed effects MMMs can be implemented in an open-source architecture, which brings substantial cost savings if it comes with a well-defined structure on its “disaggregated” modelling data. In this paper, we provide a mathematical overview of how we represent this data in a way that incorporates all of the defining business features of mixed effects MMMs. Next, we demonstrate, with a real use case, the drastic differences in insights a mixed effects model on geography can easily bring to an advertising budget in the scale of hundreds of millions of dollars per year.

1. Introduction

The marketing mix refers to variables that a marketing manager can control to influence a brand’s key performance indicators (e.g. sales, awareness, etc.). How does a key performance indicator (KPI) of interest respond to prior levels of expenditures in the marketing mix? For over 40 years, market response research has produced econometrics and time series analysis-based generalisations about the effects of marketing mix variables on different KPIs [1]. With the ever-increasing availability of data through automated feeds, use of Marketing Mix Models based on this data has increased [2]. Thus, a substantial set of end users have been using such models of the marketing mix response as an analytical input in their quest to learn from the past, optimise their future media budgets and allocate these budgets into the most profitable marketing and media channels. Such models are often named as Marketing Mix Models (MMMs) [3].

MMMs incorporate numerous factors on the nature of advertising. These include current effects, carryovers, distributed lags, saturation and competition [4]. The remaining major dimensions of advertising that a manager needs to capture (geography/market, creative, campaign messaging, product to be advertised, and sales channel) involve changes in the responsiveness of advertising exposure itself. Mixed effects models (or hierarchical linear models, without loss of generality) inherently account for the fact that model coefficients may vary between these different dimensions [5, 6, 7, 8] in addition to all the other effects (carryovers, lags, and so on). Mixed effects models also allow parameter estimation of advertising effects in dimensional combinations with very few observations; and even where data is missing in some dimensional combinations [9].

As markets globalise, marketing instruments diversify, and as technology allows customisation of marketing creatives, messages, and even brands [10], it is the norm that vast disaggregated marketing response data follows any contemporary advertising campaign. It is not uncommon to see advertisers advertise their vastly different categories of products using a large number of copy content in campaigns of varying duration, (say, two to eight weeks) to audiences across a wide variety of geographies. A model that fails to capture this heterogeneity in marketing response inherently following from marketing activity may easily bring misleading interpretations.

Historically, setting up large mixed effects models has been a complex, one-off exercise where inferring the base coefficients and variation of coefficients with dimensions involved a mathematically intricate process. Therefore, advertisers, media companies, and large agencies have relied on commercial solutions (the most popular being PROC MIXED in SAS) [11] that need specialised talent, cost substantial licensing fees, bring vendor-related limitations on the model, and yet incur additional costs for version controlling on analyses. An open-source architecture evidently eliminates licensing costs and vendor-related limitations, and potentially complements the mathematical implementation [12]. An accessible user interface can further reduce the need for users with an advanced skillset in open-source environments.

Successful implementation however, critically depends on a mathematical structure on the data representation scalable to any number of dimensions, along with built-in capabilities to account for current effects, carryovers, distributed lags, saturation and competition. A viable mathematical structure can easily trickle down to streamline the budget optimisation problem.

In this manuscript, we have put together a general mathematical framework of the modelling data powering a mixed effects MMM that embraces all major defining features of the business problem: current effects, carryovers, distributed lags, saturation, competition, and multidimensional heterogeneities. As such, the data representation is easily implementable in an open-source infrastructure and scalable to any number of features and dimensions.

The goals of this paper are twofold. First, we provide a mathematical overview of how we represent the data for mixed effects MMM in a way that incorporates all of the defining business features of mixed effects MMMs. Then, we demonstrate on a real case, the drastic differences in insights a mixed effects MMM, as simple as one on geography, can bring to an advertising budget in the scale of hundreds of millions of dollars per year.

2. Mathematical Overview of a Mixed Effects MMM

The equation (1) below describes a high-level mathematical abstraction of a mixed effects MMM:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{Z}, \boldsymbol{\xi})\boldsymbol{\beta} + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{\xi}})\boldsymbol{\gamma} + \boldsymbol{\epsilon} \qquad (1)$$

Here, $\boldsymbol{Y}$ represents the $n{\times}1$ marketing response (e.g., sales volume) vector and $\boldsymbol{Z}$ represents the $n{\times}(r+1)$ independent variable (e.g., TV gross rating point (GRP), economic indicators, weather, etc.) matrix. Specifically, $Z_{i,k}$ represents the value of the k^{th} independent variable (k can take values from 1 to $r+1$) at observation index i (i can take values from 1 to n). Every element on the first column of $\boldsymbol{Z}$ is equal to 1 to account for the intercept. $\boldsymbol{\beta}$ represents a $(r+1){\times}\,1$ dimensional coefficient vector, one for each independent variable and one for the intercept. $\boldsymbol{\gamma}$ represents a $p{\times}\,1$ multidimensional-effects-on-coefficients vector where p equals the total number of multidimensional combinations, m, times $(r+1)$. Note that when m is one, the mixed effects MMM is equivalent to a model without mixed effects. In addition, $\boldsymbol{\epsilon}$ represents the $n{\times}1$ error vector, $\boldsymbol{\xi}$ represents the $4{\times}(r+1)$ attribute matrix to account for saturation, decay, and lead/lag effects specific to each independent variable. Finally, $\boldsymbol{f}(\cdot)$ represents the $n{\times}(r+1)$ dimensional function whose elements operate on $\boldsymbol{Z}$ and $\boldsymbol{\xi}$.

The variables in the second term with the tilde mark $(\tilde{\square})$ are designed to capture the underlying multidimensional structure of the data in the model, beyond what is available in ordinary linear regression. They are a replicated and rearranged version of their original variables insofar their values are related. In particular, $\boldsymbol{f}(\boldsymbol{Z},\boldsymbol{\xi})$ relates to all multidimensional combinations as some $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{\xi}})\boldsymbol{K}$ where $\boldsymbol{K}$ is a $p{\times}(r+1)$ dimensional binary matrix (i.e., every element is either 0 or 1) and $\tilde{\boldsymbol{f}}(\cdot)$ is a $n{\times}p$ dimensional function whose elements operate on $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{\xi}}$. In line with how $\boldsymbol{f}(\cdot)$ maps to $\tilde{\boldsymbol{f}}(\cdot)$, $\tilde{\boldsymbol{Z}}\boldsymbol{K} = \mathbf{Z}$, and $\tilde{\boldsymbol{\xi}}\boldsymbol{K} = \boldsymbol{\xi}$.

Intuitively, each element of $\boldsymbol{Y}$, or Y_i is the dependent variable in one observation. The observation index i embodies both the timestamp ρ_i and the multidimensional combination μ_i (out of m such combinations). We define ρ_i in a way that resets to 1 every time index i moves from one experiment to the next. Thus, while i represents the observation index as well as the true time index when $m=1$, ρ_i always represents the time index. Due to the time series nature of the relationship (that could include both lag/lead and carryover effects), independent variables associated with this particular dependent variable are potentially located anywhere in $\mathbf{Z}$.

A "dimension" is a feature or aspect of the data which describes the level of disaggregation of the observation. We are naturally interested in testing the effects of dimensions on different independent variables. For example, geography tells us about the dependency of the incremental media effects on location. For example, we may be interested in the way in which TV GRPs affect the KPI of interest differently for each location. For this purpose, we must consider the mixed effect of geography for TV GRPs. For convenience, we henceforth assume that we have variation on just one analysis dimension (e.g. geography). Thus, m equals the number of geographies advertised, and we are going to consider the random effect for all independent dimensions[1]. This example easily extends to the general case in which we may have different multidimensional combinations of random effects for different independent variables. In practice, Equation (1) may be rewritten as:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{f}(\boldsymbol{Z},\boldsymbol{\xi})\boldsymbol{\beta} + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{\xi}})\boldsymbol{\gamma}, \qquad (2)$$

where $\widehat{\boldsymbol{Y}}$ is the estimator of $\boldsymbol{Y}$. One can further expand $\widehat{\boldsymbol{Y}}$ as shown below:

$$\hat{Y}_i = \sum_{k=1}^{r+1} \beta_k f_{i,k}(\mathbf{Z},\boldsymbol{\xi}) + \sum_{j=1}^{m(r+1)} \gamma_j \tilde{f}_{i,j}(\tilde{\mathbf{Z}},\tilde{\boldsymbol{\xi}}), \quad (3)$$

where $f_{i,k}(\cdot)$ and $\tilde{f}_{i,j}(\cdot)$ are given by

$$f_{i,k}(\boldsymbol{Z},\boldsymbol{\xi}) = \begin{cases} 1, & if\ k = 1 \\ \sum_{l=0}^{\rho_i-\xi_{1,k}-1} \hat{f}\left(\xi_{3,k},\xi_{4,k},Z_{i-\xi_{1,k}-l,k}\right)\xi_{2,k}^{l}, & otherwise. \end{cases} \quad (4)$$

[1] If random effects were not present for some independent dimensions, the entries of $\boldsymbol{\gamma}$ can be forced to zero.

$$\tilde{f}_{i,j}(\tilde{\mathbf{Z}},\tilde{\boldsymbol{\xi}}) = \begin{cases} 1, & if\ j \equiv \mu_i\ mod\ m, 1 \le j \le m \\ \displaystyle\sum_{l=0}^{\rho_i-\tilde{\xi}_{1,j}-1} \hat{f}\left(\tilde{\xi}_{3,j},\tilde{\xi}_{4,j},\tilde{Z}_{i-\tilde{\xi}_{1,j}-l,j}\right)\tilde{\xi}_{2,j}^{l}, & if\ j \equiv \mu_i\ mod\ m, m < j \\ 0, & otherwise. \end{cases} \qquad (5)$$

Here, $\xi_{1,k}$ and $\xi_{2,k}$ represent the lag/lead and carry-over parameters for an independent variable k, respectively. Having lag ($\xi_{1,k} > 0$) means that for each moment of time we must use the data from some time before (equivalent to the lag value). Leads ways in the opposite direction.

Moreover, $\hat{f}(\cdot)$ is a scalar function (e.g. exponentiation, linear, etc.) with parameters $\xi_{3,k}$ and $\xi_{4,k}$ operating on elements of $\boldsymbol{Z}$. In the case of geography as the only mixed effect dimension of interest acting on all independent variables, $\boldsymbol{\gamma}$ is a vector with the element in the j^{th} row (where j can take values from 1 to $p = m(r+1)$) equal to random effect associated with the $\left\lceil\frac{j}{m}\right\rceil$-indexed independent variable and $(j-1)\ mod\ (m+1)$-indexed geography. Thus, the vector $\boldsymbol{\gamma}$ only has m independent entries per variable plus m intercept (hence size of $m(r+1)$). Equation (3) as described above is valid for all observation indices that has the investment on the time period belonging to this index counted at least once in the model. In mathematical terms, we look for an observation index i to conform to the condition $\mathcal{I}_i$:

$$\mathcal{I}_i = \{i | I_{min_i} \le i \le I_{max_i}\}, \qquad (6)$$

where

$$\begin{aligned} I_{min_i} &= s(i) + \underset{\square}{\max}\left(0, \max_k \xi_{1,k}\right) \\ I_{max_i} &= e(i) + \underset{\square}{\min}\left(0, \min_k \xi_{1,k}\right). \end{aligned} \qquad (7)$$

Here, $s(i)$ and $e(i)$ represents the starting and ending time indices for the geography to which observation index i belongs: ρ_i and μ_i denotes the time and geography indices respectively. Moreover:

$$\mu_i = \left\lfloor\frac{i-1}{\left\lfloor\frac{n}{m}\right\rfloor}\right\rfloor + 1, \qquad (8)$$

$$\rho_i = i - \left\lfloor\frac{n}{m}\right\rfloor(\mu_i - 1), \qquad (9)$$

$$s(i) = \left\lfloor\frac{n}{m}\right\rfloor(\mu_i - 1) + 1, \qquad (10)$$

$$e(i) = \left\lfloor\frac{n}{m}\right\rfloor\mu_i. \qquad (11)$$

Finally, the $n \times m(r+1)$ dimensional $\tilde{\boldsymbol{Z}}$ and the $4 \times m(r+1)$ dimensional $\tilde{\boldsymbol{\xi}}$ can be explicitly defined in terms of $\boldsymbol{Z}$ and $\boldsymbol{\xi}$ as shown below:

$$\tilde{Z}_{i,j} = \begin{cases} Z_{i,\frac{j-\mu_i}{m}+1}, & if\ j \equiv \mu_i\ mod\ m \\ 0, & otherwise. \end{cases} \qquad (12)$$

and

$$\tilde{\xi}_{i,j} = \xi_{i,\left\lfloor\frac{j-1}{m}\right\rfloor+1}. \qquad (13)$$

We should note that prior to inferring MMM parameters, a modeller might want to undertake a set of pre-processing steps and decisions in order to formulate a more representative model. First, the functional form of $\hat{f}(\cdot)$ can be chosen from a set of alternatives. In our implementation, we typically use variants of the functional forms shown in Table 1. The modeller may also choose to ensure dependent variables or independent variables (i.e. Y_i or $Z_{i,k}$, $\forall k > 1$ and $\forall i$) are normalised so that they have a similar mean within different geographies across time.

Table 1. Our functional form choices for implementing saturation of independent variables.

Name	$\hat{f}(\xi_3, \xi_4, Z) =$
Linear	Z
Logarithmic	$\ln\left(\underset{\square}{\max}(Z,1)\right)$
Power	Z^{ξ_3}
Exponential	$1 - e^{-\frac{Z}{\xi_3}}$
S-shaped	$\frac{\xi_4}{10^{10}}\xi_3^{100Z/\underset{\square}{\max} Z}$

2.1. Extension to Multidimensional Random Effects

Dimensions are the features which describe the level of disaggregation of the collected data. We usually consider five dimensions: geography, creative, campaign, product, and outlet:

$$\boldsymbol{G} = \{geography, creative, campaign, product, outlet\}. \tag{14}$$

The specific structure of matrix $\boldsymbol{Z}$, vector $\tilde{\boldsymbol{\xi}}$ and function $\tilde{\boldsymbol{f}}$ as described above are all dependent on a number of multidimensional combinations (or geographies present in the single dimensional case). The Online Appendix provides an illustration of this dependency for the case of multiple experiments on a single dimension accounting for all independent variables.

In a more general case, any desired subset of multidimensional combinations can be applicable for each individual variable. We adopt a modified Cartesian approach to position data to extend our formulation. Let $\boldsymbol{G}_k$ be the selected subset for each variable k, i.e. $\boldsymbol{G}_k \subseteq \boldsymbol{G}\ \forall k$. For each variable, zero, one or more dimensions may be selected to consider random effects. For example, assume that in addition to m_1 geographic regions, we also want to do the experiment on different products with m_2 variations. Each m_q represents the number of factors for each individual dimension (e.g. geography, product, outlet, etc.). Therefore, the number of multidimensional combinations for variable k, $M_k$becomes:

$$M_k = \prod\nolimits_{q \in \boldsymbol{G}_k} m_q. \tag{15}$$

The $\boldsymbol{\gamma}$ vector is divided into multiple subsets corresponding to the intercept and independent variable. The length of the corresponding block in $\boldsymbol{\gamma}$ for each variable is equal to its total number of mixed effects coefficients. For the intercept, number of elements in $\boldsymbol{\gamma}$ will be:

$$M_0 = \prod\nolimits_{q \in \boldsymbol{G}_0} m_q, \tag{16}$$

where

$$\boldsymbol{G}_0 = \cup_{k=1}^{r+1} \boldsymbol{G}_k \tag{17}$$

Therefore, we apply a naïve Cartesian representation with the only distinction of choosing not to add rows to the mixed effects vector for dimensions that are not applicable for a particular variable[2], to ensure full identifiability of mixed effects (i.e., no variable can be unique to a dimension). Therefore, some independent variables might have higher dimensional combinations for its mixed effects and some variables not. Expressed in this way, the length of $\boldsymbol{\gamma}$ becomes:

$$p = \sum\nolimits_{k=0}^{r} M_k. \tag{18}$$

2.2. Implementation

A commonly used way to infer mixed effects MMM is simultaneous use of exhaustive search within commercial regression solvers (such as PROC REG and PROC MIXED in SAS). However, these approaches can mostly handle up to single dimensional mixed effects (e.g., only based on geography). With our data representation, we were able to migrate our process into the open-source language R (R Core Team, 2013), where the *lme4* library (Bates, Mächler, Bolker, & Walker, 2015) is capable of handling multiple dimensions with an arbitrary combination for each independent variable. This language also gives us the flexibility to employ distributable search methods (e.g., exhaustive grid search) to search for the non-coefficient parameters of the mixed effects MMM (e.g., $\boldsymbol{\xi}$) that provides the best fit.

2.3. Optimising MMM Output

The natural next step after creating the best model for the observed data is to use that model and optimise future cross-channel media investments. Once all model parameters are

[2] We may still force the relevant elements of $\boldsymbol{\gamma}$ to zero if some level of a particular dimension is not applicable to the variable (but some other levels of this dimension are).

either optimised or fixed and therefore a complete model relating the independent variables to the dependent variable of interest, $\boldsymbol{Y}$ is already built, the optimisation problem can be stated as:

$$\begin{aligned} \boldsymbol{Z}^* &= \underset{\boldsymbol{Z}}{\operatorname{argmax}} \sum_{i=1} \widehat{\boldsymbol{Y}}_i(\boldsymbol{Z}) \\ &\mathit{subject\ to} \sum_{i=1}^{n}\sum_{k=1}^{r+1} \eta_{i,k} Z_{i,k} \leq I \\ &\boldsymbol{Z} \in [\boldsymbol{Z}_L, \boldsymbol{Z}_U] \end{aligned} \qquad (19)$$

where I is the total available investment and $\boldsymbol{\eta}$ is the investment cost per unit of k^{th} variable, where the intercept has no cost; i.e., $\eta_{\cdot,1} = 0$. We denote lower and upper quantity limits on marketing variables as $\boldsymbol{Z}_L$ and $\boldsymbol{Z}_U$. One can also put weights on each observation index designed to capture aspects of marketing investment that are not included in this mixed effects MMM such as seasonality effects (e.g., some media channels have higher relative impact at particular times of the year), or discounting factors for longer term planning. This approach allows non-investment variables (e.g., weather, macroeconomic indicators, etc.) to influence the resulting optimal investment mix.

A variety of methods can be employed to achieve the optimal solution, the easiest ones are steepest descent and gradient-based Newton search (multi-start variants if the objective function does not happen to exhibit diminishing returns). One obvious drawback with such general-purpose strategies is loss of efficiency and unnecessary dependence on the numerical properties of the objective function, sometimes well off the optimal solution.

3. Case Study

We demonstrate the potential impact of mixed effect MMMs on a historical dataset of a premium segment automobile brand, where the KPI of interest is sales volume. To model responsiveness of the marketing mix to the KPI, we consider MMMs both with and without mixed effects. The underlying time series data spanning three and a half years includes weekly information broken down by geography on: (i) sales, (ii) marketing spend of the brand in different channels, (iii) marketing spend of competitors, (iv) macro-economic variables and (v) weather patterns. Using this data, we present the changes in the estimated coefficients of different marketing channels and investigate the changes in the magnitude of coefficients across different geographies. Finally, we discuss the importance of the results on budget allocation.

Table 2. Categorised summary of variables present in the dataset.

Category	Count	Sample Variable
Economy	44	Dow Jones Industrial Average
GRP	6	National TV GRPS AD3554
Holiday	75	Good Friday Week After
Marketing	85	Display Impressions
KPI	6	Industry Total Autodata Sales
Spend	32	Display Spend
Weather	21	Max Temperature
Other	22	Industry Total Autodata Incentives

3.1. Dataset

The dataset includes weekly data of a premium segment automobile brand in all 210 Designated Market Areas (DMAs) in the United States starting from the week of January 1st, 2012 until the end of the week of July 27th, 2015. We hold-out the last 26 observations for testing the predictions of the model and use the rest of the 161 observations for building models. The dataset includes 32,949,917 records related to 291 variables, coming from 210 DMAs (210 geographies), 106 different products of the same company and competitors, 802 campaigns on 18 outlets, and with 1645 registered different creatives, although not all combinations exist. The categorisation of variables and a sample variable in each category is presented in **Table 2.**

Figure 1 presents changes in U.S. nationwide sales of the target product over time. Sales show a regular pattern with expected seasonal changes.

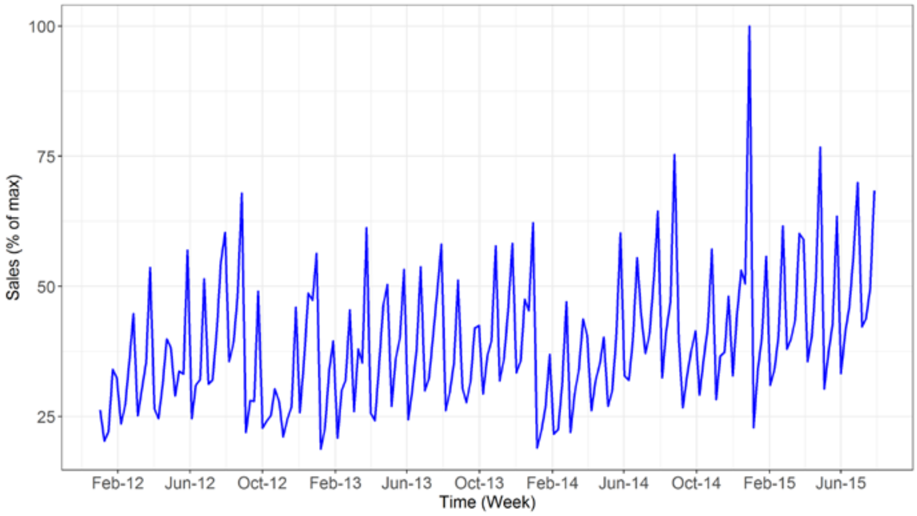


Figure 1. Changes in target product sales during the modelling time period

In **Figure 2**, we show the changes in media spending on different selected advertising activities over time.

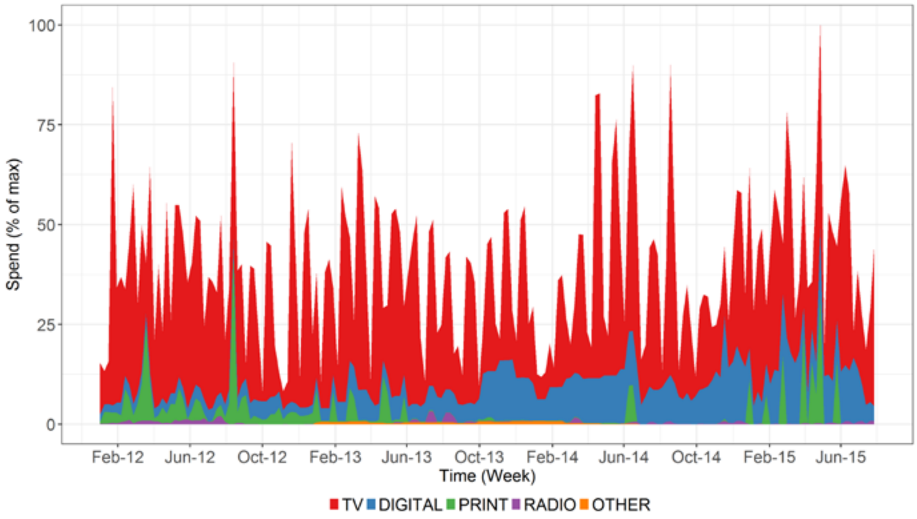


Figure 2. Selected media spending on different channels during the modelling time period

We can see the changes in Gross Rating Points (GRPs) in variables that are measured by GRPs in **Figure 3**.

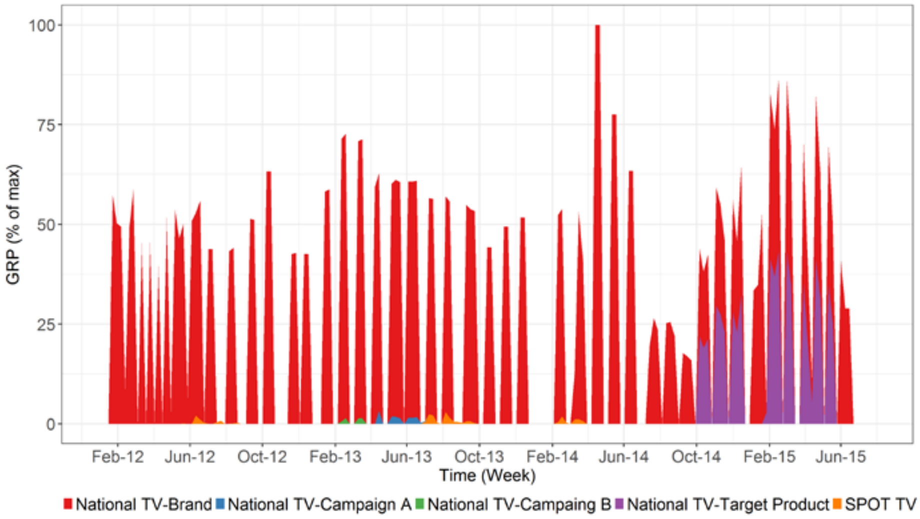


Figure 3. Changes in Gross Rating Points (GRPs) acquired through the modelling period

3.2. Results

To illustrate how introducing mixed effects on geography affects insights we infer from marketing mix models, we built two separate models with the same underlying data and variables. In the first model, we do not consider any mixed effects; while in the second model we introduce mixed effects on geography across three variables. The three variables where we apply geography mixed effects from each DMA are: (1) National TV spend, (2) Local TV spend, and (3) magazine spend. We can describe the quality of fit of the two models within the model data as in **Table 3**.

As Table 3 suggests, the predictive accuracy of the mixed effects MMM is much better within sample. To see whether this increase in predictive performance simply follows from overfitting, we forecast the U.S. nationwide sales on the hold-out dataset of 26 weeks following the modelling period. We can summarise the predictive performance of the two models on the hold-out data as in **Table 4**.

Inclusion of mixed effects genuinely improve summary measures of bias and accuracy. **Figure 4** illustrates the out-of-sample forecasts for an MMM without mixed effects and **Figure 5** shows the improvement in forecasts when we include mixed effects on geography to TV and magazine spend.

Table 3. Comparison of the in sample fit of the two models

Measure	No Mixed Effects	Mixed Effects
R^2	0.91	0.95
Adjusted R^2	0.91	0.95
Durbin-Watson statistic	1.53	1.76
Intercept Coefficient	-1.27	-0.84
Intercept t-Value	-59.20	-19.52

Table 4. Out of sample performance of the two models

Measure	No Mixed Effects	Mixed Effects
Mean Error	-4.62	-3.12
Root MSE	96.50	55.64
Mean Absolute Error	78.46	45.35
Mean Percentage Error	-4.83	-2.74
Mean Absolute Percentage Error	21.20	11.74

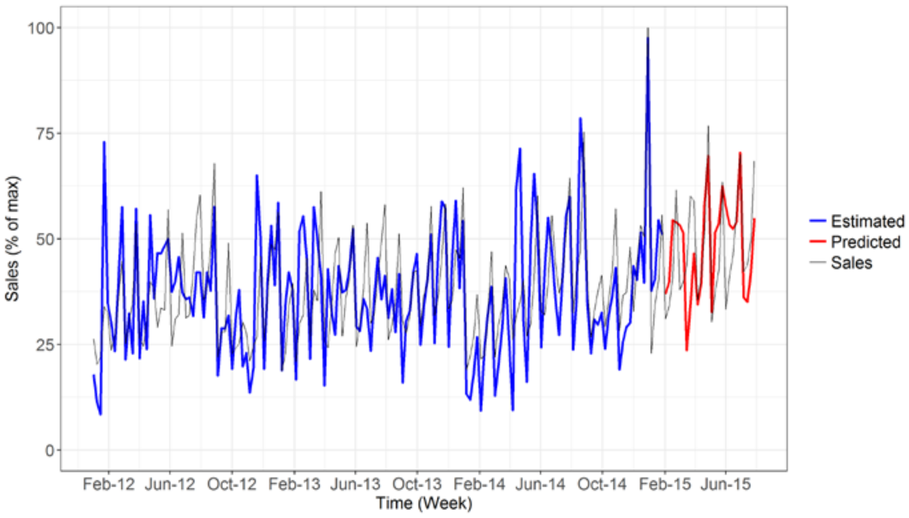


Figure 4. Forecasting performance of MMM without mixed effects on the 26-week hold-out data

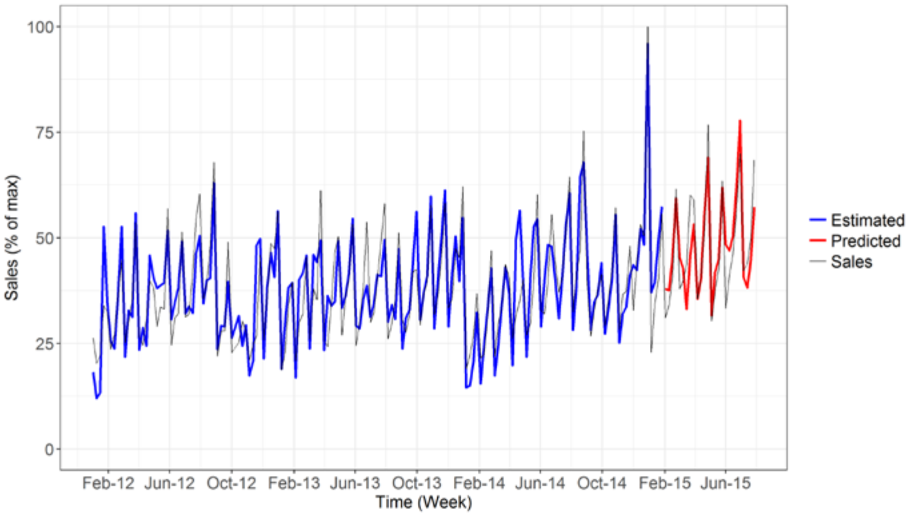


Figure 5. Forecasting performance of the mixed effects MMM on the 26-week hold-out data

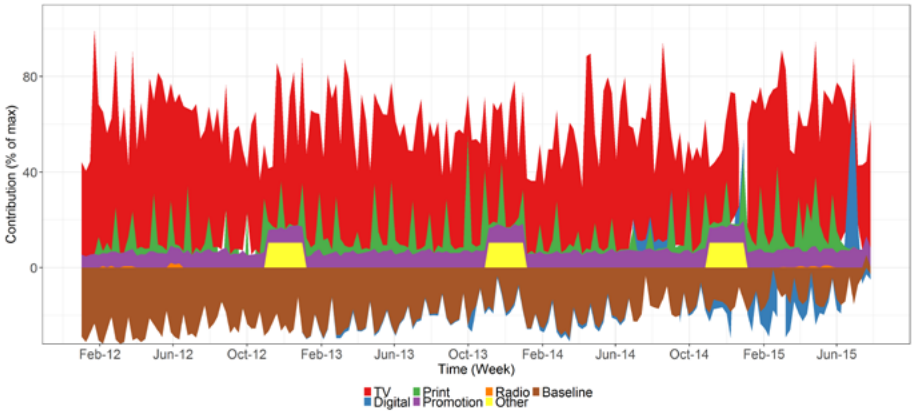


Figure 6. Contribution of media variables to spending based on MMM without mixed effects

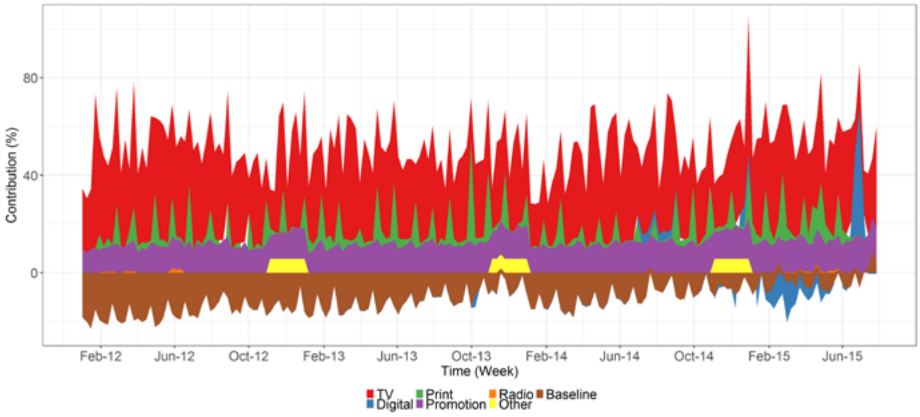


Figure 7. Contribution of media variables to spending based on mixed effects MMM

3.3. Discussion

Insights generated by MMMs are used to understand the performance of different media outlets as well as guiding the budget allocation practice. Therefore, having a more accurate model provides an extra benefit of improved budget allocation in subsequent campaigns. Here we discuss how contrasts in the two models translates into new business insights.

We depict the contribution of variables to the total sales on an MMM without random effects, and on a mixed effects MMM in **Figures 6** and **7**, respectively.

We notice that TV spend systematically gets a lower contribution when mixed effects are included. This can help the planners shift investment out of expensive TV advertising. Magazines perform similarly and digital display performs slightly better.

Furthermore, changes in the coefficients of the two models across geographies has interesting interpretations for the decision makers. For example, we observe a better marginal performance for local TV in lifting sales across DMAs with lower sales. **Figure 8** presents geographical impact of local TV compared to the number of sales at each DMA on the U.S. map.

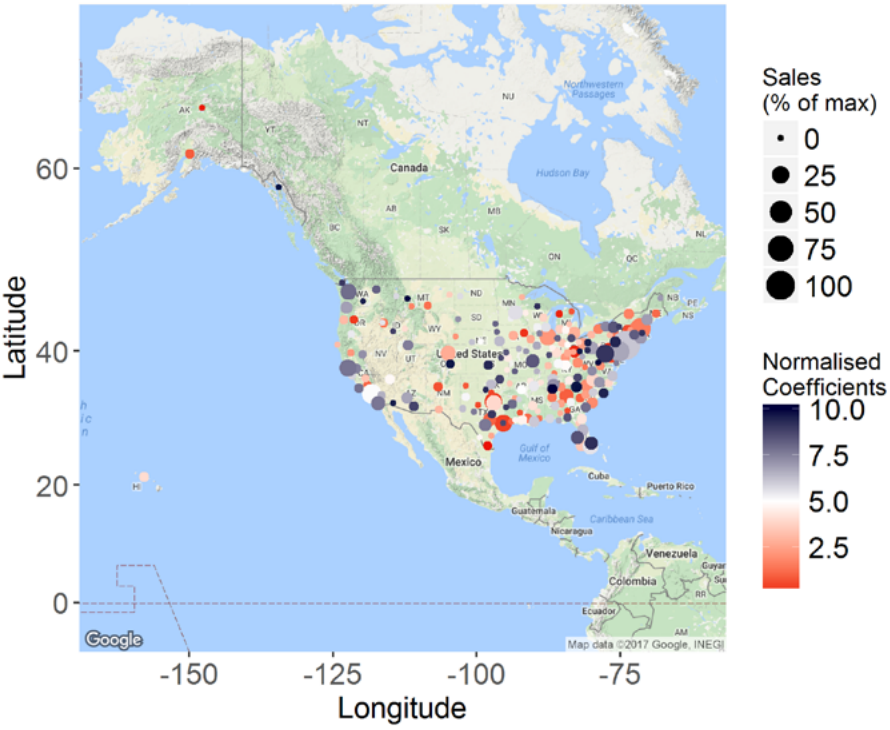


Figure 8. The estimated coefficient for total spending on local TV across different geographies

Similarly, spend on the magazines shows a similar relationship with sales. The larger the number of sales is in a geographical area, the lower the effect of the magazine spending becomes, which highlights the impact of magazines in harder-to-target geographical areas (**Figure 9**).

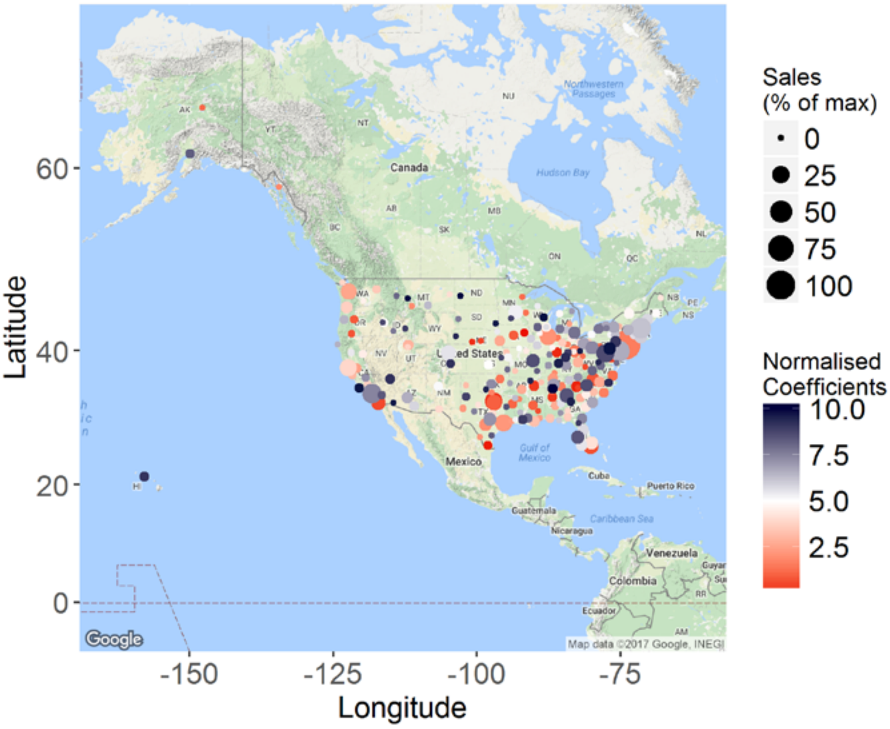


Figure 9. Estimated coefficient for magazine spending across different geographies

Conversely, we observe no meaningful pattern on national TV coefficients regarding the number of sales in each geography. This result is expected considering the broadcast nature of national TV. We illustrate the lack of relationship in **Figure 10.**

With a mixed effects MMM, we achieve higher confidence to cut back on TV spending, to consider local TV spending over national only in areas with low sales, and to focus on magazines in harder-to-target geographical areas. Given that the total TV spend easily exceeds a million dollars each week, potential for substantial savings are abound.

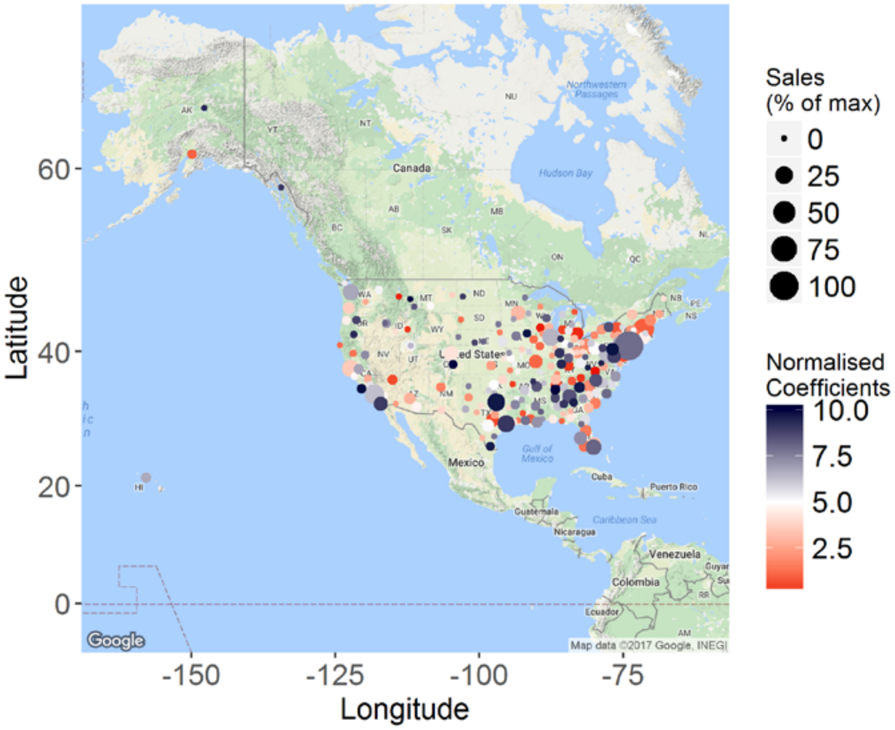


Figure 10. Estimated coefficient for total spending on national TV across different geographies

4. Conclusions

How do agencies incorporate the heterogeneity in the response to media across geographies and other dimensions? Mixed effects MMMs helps quantify this challenge in a systematic and effective manner. In this paper, we provide a mathematical overview of how data in a mixed effects MMM can be represented in a way that incorporates all of the defining business features of MMMs such as carryovers, lags, and saturation in addition to variations in responsiveness to KPIs across markets and creatives. Through this structure, we are able to implement mixed effects models in an open-source architecture and freely scale it to multidimensional effects without difficulty. We also demonstrate on a real case, with advertising spend well in the order of millions of dollars per year, that we can easily achieve substantial differences in insights through a mixed effects model.

Reducing analyst time for statistical processing and scenario analysis is paramount. The budget optimisation problem might exhibit several attractive properties (e.g. polynomial ratios, additive separability, etc.). We can exploit these properties to develop tailored and hence, very efficient, optimisation routines enabling comprehensive scenario analyses. Our experience in implementation tells us that applying mixed effect MMMs throughout the modelling process guides model building efforts hand in hand with exploratory data analysis to ensure the model input possesses necessary statistical quality (e.g., mixed effects and regression coefficients are identifiable, multicollinearity is not present, transformed media inputs exhibit sufficient variability).

Acknowledgments

We would like to appreciate the kind help of Mr. Jeremy Rose, Mr. Timothy Collins, Mr. Ikechi Okoronkwo, and Mr. Jacob Wan in providing the dataset for this analysis.

References

1. D. M. Hanssens, L. J. Parsons and R. L. Schultz, Market Response Models: Econometric and Time Series Analysis., New York, NY: Kluwer Academic Publishers, 2001.

2. S. Gupta and T. J. Steenburgh, Allocating Marketing Resources, Cambridge, MA: Harvard Business School, 2008.

3. W. A. Cook and V. S. Talluri, "How the Pursuit of ROMI Is Changing Marketing Management," Journal of Advertising Research, vol. 44, no. 3, pp. 244-254, 2004.

4. G. J. Tellis, "Modeling Marketing Mix," in The Handbook of Marketing Research: Uses, Misuses, and Future Advances, Thousand Oaks, CA, Sage Publications, Inc., 2006, pp. 506-522.

5. M. J. Lindstrom and D. M. Bates, "Mixed Effects Models for Repeated Measures Data," Biometrics, vol. 46, no. 3, pp. 673-687, 1990.

6. G. J. Tellis, R. J. Chandy and P. Thaivanich, "Decomposing the e ects of direct advertising: Which brand works, when, where, and how long?," Journal of Marketing Research, vol. 37, no. 1, pp. 32-46, 2000.

7. R. J. Chandy, G. J. Tellis, D. J. Macinnis and P. Thaivanich, "What to say when: Advertising appeals in evolving markets," Journal of Marketing Research, vol. 38, no. 4, pp. 399-414, 2001.

8. D. M. Hanssens, K. H. Pauwels, S. Srinivasan, M. Vanhuele and G. Yildirim, "Consumer attitude metrics for guiding marketing mix decisions," Marketing Science, vol. 33, no. 4, pp. 534-550, 2014.

9. P. Bhattacharya, "Marketing Mix Modeling: Techniques and Challenges," in NCSU SESUG Proceedings, St. Pete Beach, FL, 2008.

10. J. P. Gownder, "Mass Customization Is (Finally) The Future Of Products," Forrester Research, Cambridge, MA, 2011.

11. M. Minelli, M. Chambers and A. Dhiraj, Big Data, Big Analytics: Emergng Business Intelligence and Analytic Trends for Today's Businesses, Hoboken, NJ: Wiley & Sons Inc., 2013.

12. R. Muenchen, R for SAS and SPSS Users, New York, NY: Springer Science+Business Media, 2011.

13. R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing, 2013.

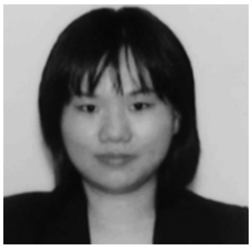
14. D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed Effects Models Using Ime4," Journal of Statistical Software, vol. 67, no. 1, pp. 1-48, 2015.

Authors

Saeed R. Bagheri, Ph.D. is currently the Director of Analytics and Insights at Amazon Advertising. Prior to joining Amazon and at the time of writing this paper, Saeed led GroupM's Global Data and Analytics Product and R&D team. There, he looked after all data and analytics related products globally from inception all the way to deployment, training and maintenance. Prior to this role, he was at Philips Research leading the Global Healthcare Services Innovation Topic as well as North America Services.

Seyed Hanif Mahboobi, Ph.D. was a Data Science Manager at GroupM at the time of the writing of this paper. He has several years of experience in high-performance computing and mathematical modelling and has developed advanced methodologies for advertising performance measurement. He has recently joined Amazon Web Services as a Data Scientist. Seyed Hanif's recent focus is building scalable AI/ML solutions using cloud computing.

Mericcan Usta, Ph.D. was a Data Scientist at GroupM at the time of the writing of this paper. Mericcan is a researcher, practitioner, and educator in Systems Engineering and Supply Chain Management. He is experienced in software applications of optimisation theory, resource allocation, statistical inference, machine learning, mathematical models of advertising response, supply chains, as well as the U.S. criminal justice system. He is currently an Operations Research/Data Scientist at Apple.



Jing Zhao, Ph.D. is currently a Senior Data Scientist at GroupM. She has years of experience building data products leveraging statistics and machine learning for media measurement and planning. Her areas of interest include marketing mix modelling, cross-channel attribution, consumer insights and audience targeting.

Hamid R. Darabi, Ph.D. is currently a Senior Data Scientist at Remedy Partners. Prior to that, he was a Post-Doctoral Research Scientist at GroupM. He has years of experience in leading, developing, and productising predictive models. His main research focus includes applying machine learning modelling techniques and optimisation algorithms in different industries such as marketing, healthcare, and transportation.

Leverage Social Media Data to Explore Fashion Trending

Ling Huang
Tumblr Inc.

Amanda Brennan
Tumblr Inc.

Classifications, Key Words:

- Social Media Analysis
- Fashion Trending
- Cluster Analysis

Abstract

Social media platforms are rapidly growing channels for users to express themselves through rich media formats including text, images, animated GIFs, music, video and more. As more people turn to social media to explore their interests, advertisers leverage data from these platforms to make their product development and marketing strategies feel more authentic. The typical product development cycle for detecting fashion trends contains four stages: discovering trends, validating them, monitoring growth, and making decisions based on the learnings. In the discovery stage, aggregating post data from major social media platforms enables us to quantify emerging trends from top fashion influencers and other aspirational fashion retailers in the market. Leveraging social media search data facilitates knowledge on growth and momentum of trends during the validation stage. In the monitoring stage, we use cluster models to generate style groupings which continuously monitor trends and help the user detect significant changes. Afterwards, we assess results perpetually, adapting at speed and amending the approaches to curate trends and applications to get better products and strategies to deliver to the consumer. Tumblr is one of the most popular social blogging platforms where users can create and share posts with followers. This paper presents innovative use cases in our pilot study to leverage social media data to uncover up-and-coming trends, generate in-depth reports and predictive models on fashion-forward millennial customers, and help business stakeholders in their decision-making.

1. Introduction

Social media platforms such as Facebook, Instagram, Tumblr, etc., allow users to create content and interact with each other. Users can express themselves through rich media formats including text, images, animation, music, video and more. The interactions among users can take many forms, but some common types include updating public profiles, posting recent activities, sharing interesting content and commenting on popular topics. Based on research by Barnes and Daubitz (2017), social networks are becoming the top choice for marketing channels, which ranked higher than online advertising, traditional print/ broadcast media, business directory listings, daily deal sites, etc. [1].

For fashion brand owners and shareholders, the typical product development process is a repetitive cycle with multiple steps including discovery, curation, editing, sending to makers, developing, planning, buying, production and transit. Each cycle could last six months to a year. The key challenge of leveraging rich social media data is to collect, explore and analyse large amounts of data to identify user needs while considering trade-offs between allocated time, devoted resources and the return on investment. Generally speaking, there are four major stages for detecting fashion trends including discovery, validation, monitoring, and learning. The detailed descriptions of each stage are listed below:

- Discovery stage is where we detect trends as they emerge through data from social media listening and measuring trending topics
- Validation stage empirically supports or refutes the detected trending styles and contributes to the story with facts and data intelligence
- Monitoring stage involves continually evaluating trends to validate decision making, providing insight to support product assortment processes and the agility to adapt up to the last possible moment
- Learning stage is when the results are assessed and where we can adapt quickly, amending the approach to curate trends and applications to get better products and strategies to the consumer

Each stage will require data analytics to address real time fashion in order to create trend driven products, and thus, increase consumer engagement. Our goal is to leverage all social media data sources to uncover up-and-coming trends, creating in-depth reports on fashion-forward customers, and help companies plan product assortments and marketing strategies.

2. Analysis on Trending Fashion

2.1. Rank Posts to Discover Fashion Trends

In order to find out what's buzzy in the fashion world, we evaluate the posts that users are making as well as their browsing habits. With brands relying heavily on social media to promote their products and services, along with market information and consumer feedback, our research goal is to use large-scale social media data to identify trending fashion styles.

Tumblr is a popular micro-blogging service platform, which has millions of users every month. A Tumblr blog contains a profile picture, blog title, and blog description appearing at the top, followed by a stream of blog posts below. Unlike other social networks, Tumblr users have the ability to completely customise their profile's look and feel through the blog network, both on the web and in the mobile app. Common user activities throughout Tumblr include: 1) creating a post on one's blog; 2) sharing a post created by someone else to a third party via a link or their internal messaging service; 3) sharing a post someone else made to their own blog through reblogging; 4) liking a post by another blog; and 5) following another blog. An example of a Tumblr post is shown in **Figure 1**.

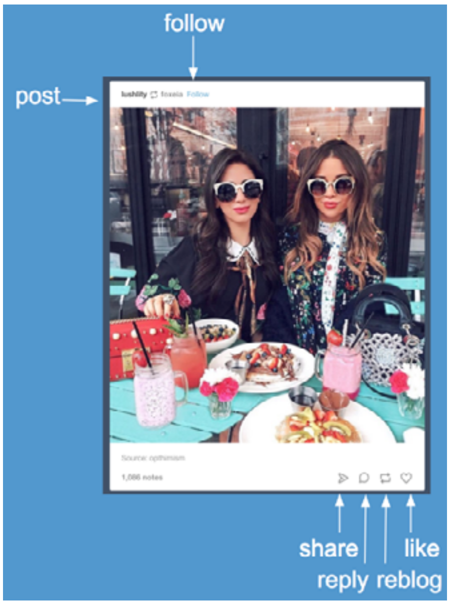
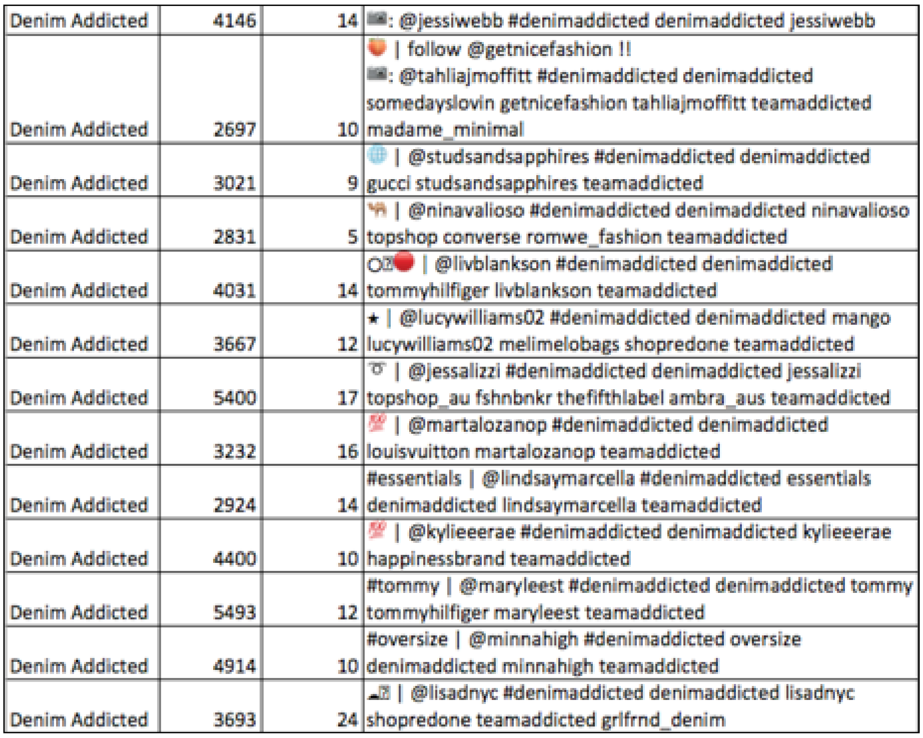


Figure 1: Example of Tumblr Post

Our system analyses not only Tumblr internal user activities, but also external data, collecting, aggregating, and summarising the data from various data sources including Instagram and Pinterest. A combination of internal data and external data is used to generate visual reports presenting top posts to provide trend inspiration insights. The posts from other platforms look similar to those made on Tumblr. The data points received from other platforms include, but are not limited to, a posted photo, related description, number of likes, and number of comments. Numerous APIs are applied to acquire post data from above social media platforms [2] [3]. See Table 1 for a sample of the metadata extracted from Instagram posts for business stakeholders. Note that users can use 'likes' and 'comments' to send explicit appreciation or endorsements of the posts. A 'reply', 'reblog' with comments, and the descriptions of the posts can be used to determine what the relevant conversations happening in the post include. From there, we can extract and measure trending styles to identify what's trending over time.

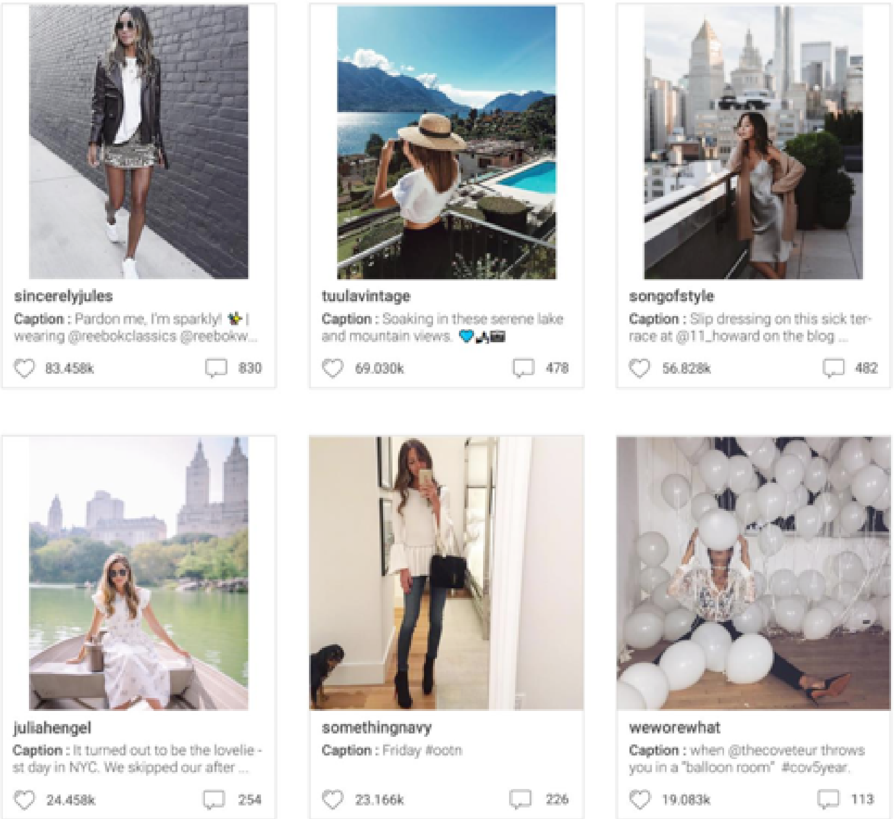
Brand engagement metrics such as likes or comments can be applied to rank all related posts. We can provide our business stakeholders these overarching reports to reveal current trending fashions to them while they are in the discovery stage. **Figure 2** shows an example of a monthly

Table 1: Example of metadata from Instagram posts in the hashtag #denimaddicted



Denim Addicted	4146	14	📷: @jessiwebb #denimaddicted denimaddicted jessiwebb
Denim Addicted	2697	10	🍑 | follow @getnicefashion !! 📷: @tahliajmoffitt #denimaddicted denimaddicted somedayslovin getnicefashion tahliajmoffitt teamaddicted madame_minimal
Denim Addicted	3021	9	🌐 | @studsandsapphires #denimaddicted denimaddicted gucci studsandsapphires teamaddicted
Denim Addicted	2831	5	🐪 | @ninavalioso #denimaddicted denimaddicted ninavalioso topshop converse romwe_fashion teamaddicted
Denim Addicted	4031	14	O?🔴 | @livblankson #denimaddicted denimaddicted tommyhilfiger livblankson teamaddicted
Denim Addicted	3667	12	★ | @lucywilliams02 #denimaddicted denimaddicted mango lucywilliams02 melimelobags shopredone teamaddicted
Denim Addicted	5400	17	♉ | @jessalizzi #denimaddicted denimaddicted jessalizzi topshop_au fshnbnkr thefifthlabel ambra_aus teamaddicted
Denim Addicted	3232	16	💯 | @martalozanop #denimaddicted denimaddicted louisvuitton martalozanop teamaddicted
Denim Addicted	2924	14	#essentials | @lindsaymarcella #denimaddicted essentials denimaddicted lindsaymarcella teamaddicted
Denim Addicted	4400	10	💯 | @kylieeerae #denimaddicted denimaddicted kylieeerae happinessbrand teamaddicted
Denim Addicted	5493	12	#tommy | @maryleest #denimaddicted denimaddicted tommy tommyhilfiger maryleest teamaddicted
Denim Addicted	4914	10	#oversize | @minnahigh #denimaddicted oversize denimaddicted minnahigh teamaddicted
Denim Addicted	3693	24	▁? | @lisadnyc #denimaddicted denimaddicted lisadnyc shopredone teamaddicted grlfrnd_denim

Figure 2: Example of trending data across social media platforms



sincerelyjules
Caption : Pardon me, I'm sparkly! ✨ | wearing @reebokclassics @reebokw...
♡ 83.458k 💬 830

tuulavintage
Caption : Soaking in these serene lake and mountain views. 💙⛰📷
♡ 69.030k 💬 478

songofstyle
Caption : Slip dressing on this sick terrace at @11_howard on the blog ...
♡ 56.828k 💬 482

juliahengel
Caption : It turned out to be the loveliest day in NYC. We skipped our after ...
♡ 24.458k 💬 254

somethingnavy
Caption : Friday #ootn
♡ 23.166k 💬 226

weworewhat
Caption : when @thecoveteur throws you in a "balloon room" #cov5year.
♡ 19.083k 💬 113

trending report for posts from Instagram. It is a fast way to learn what fashion tags are steering the conversation in order to assess what that means for the business.

2.2. Rank Keywords to Validate Fashion Trends

Fashion trends and memes happen fast across platforms. Understanding how style preferences change with time is of critical importance to a fashion retailer, especially in the validation stage. Consumers will use keywords across their social media posts ranging from colour, print, fabric, silhouette and other details to describe the fashion trends they are posting about. Identifying and anticipating which, if any, of these trends could affect user preferences would enable more effective control of a retailer's inventory. It is also important for the retailer to understand what language consumers are using to describe popular trends, so they can adapt their products or marketing language to be more aligned with the consumers' desires. This will help promote the retailers' business, showing they are familiar with what is popular with those who show affinity for the brands or styles they offer.

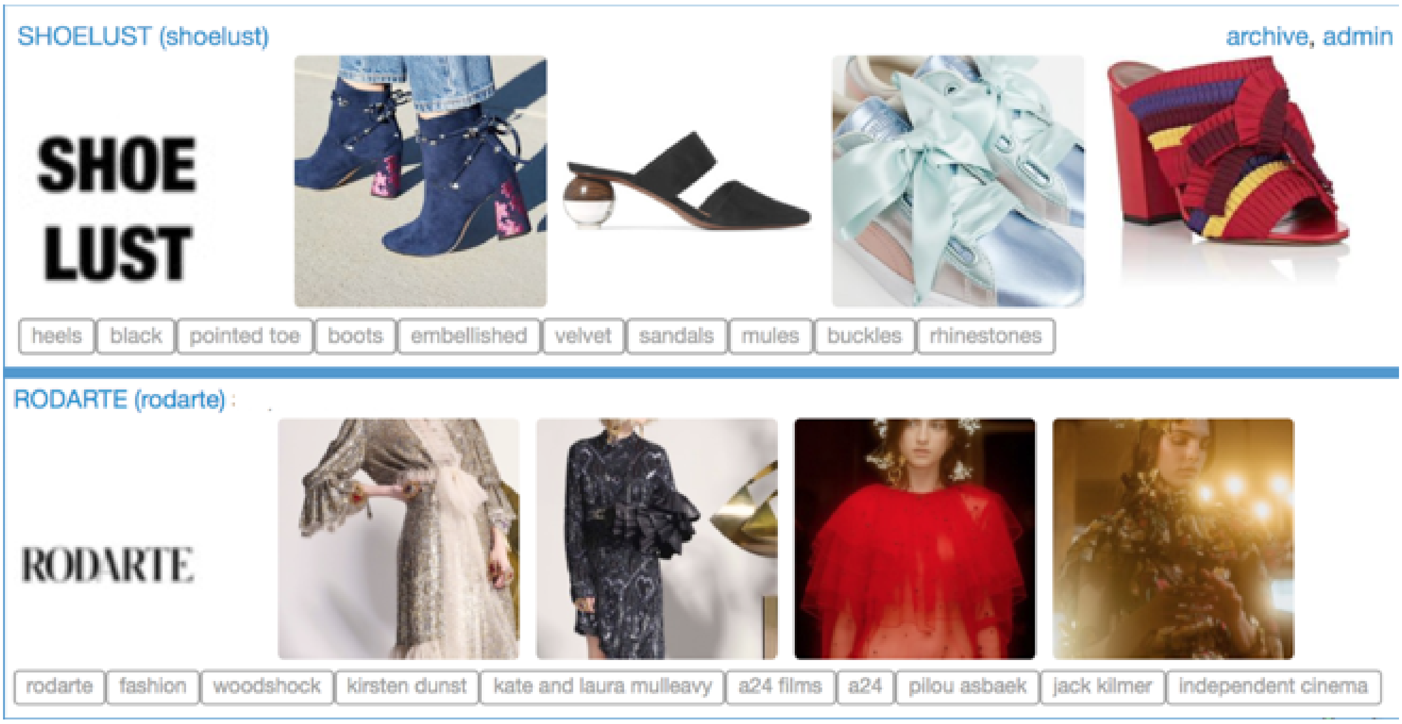
Tumblr represents a unique combination of rich and diverse content in a dynamic social network. In order to obtain useful data on trending styles, our system extracts keywords from available blog data, including user post descriptions and engaged messages. In addition, any post type can be annotated by a user with words starting with the "#" sign (called tags) that concisely describe a post and allow for easier browsing and searching. An example of tagging information is shown in **Figure 3**.

Each keyword, or tag, might have a distinct meaning when it comes to trending fashion styles. Thus our system analyses keywords by investigating three metrics:

- looking at keyword volume for popularity
- looking at keyword growth rates to see which trends are growing or declining
- keyword momentum, or the rolling moving average that signals the direction of a trend, growing or declining

Accordingly, this rank is defined as a scaled search index over the last 6 months to indicate keyword popularity; growth is defined as the last 6-month total search this year (TY) compared to same period last year (LY) to indicate long-term patterns; and momentum is last month's trend compared to a 3-month trend to show short-

Figure 3: Example of Tumblr tags on fashion blogs



term patterns. All metrics should be refreshed monthly to capture short-term and long-term changes.

For comparison purpose, all metrics are scaled to range between 0 and 100. **Figure 4** shows an example of a trend report from January 2017. In this report, sweatshirts rank highest among top trends, having 33% growth year over year (YoY) and relatively high momentum at 12%. Bomber jackets have 158% growth YoY and very high momentum at 19%. Though the term is searched for less often, track pants are performing better than joggers, with 12% growth vs -1% growth. Accordingly, we would recommend a retailer refresh their product assortment with more sweatshirts, bomber jackets and track pants for fashion-forward millennial customers.

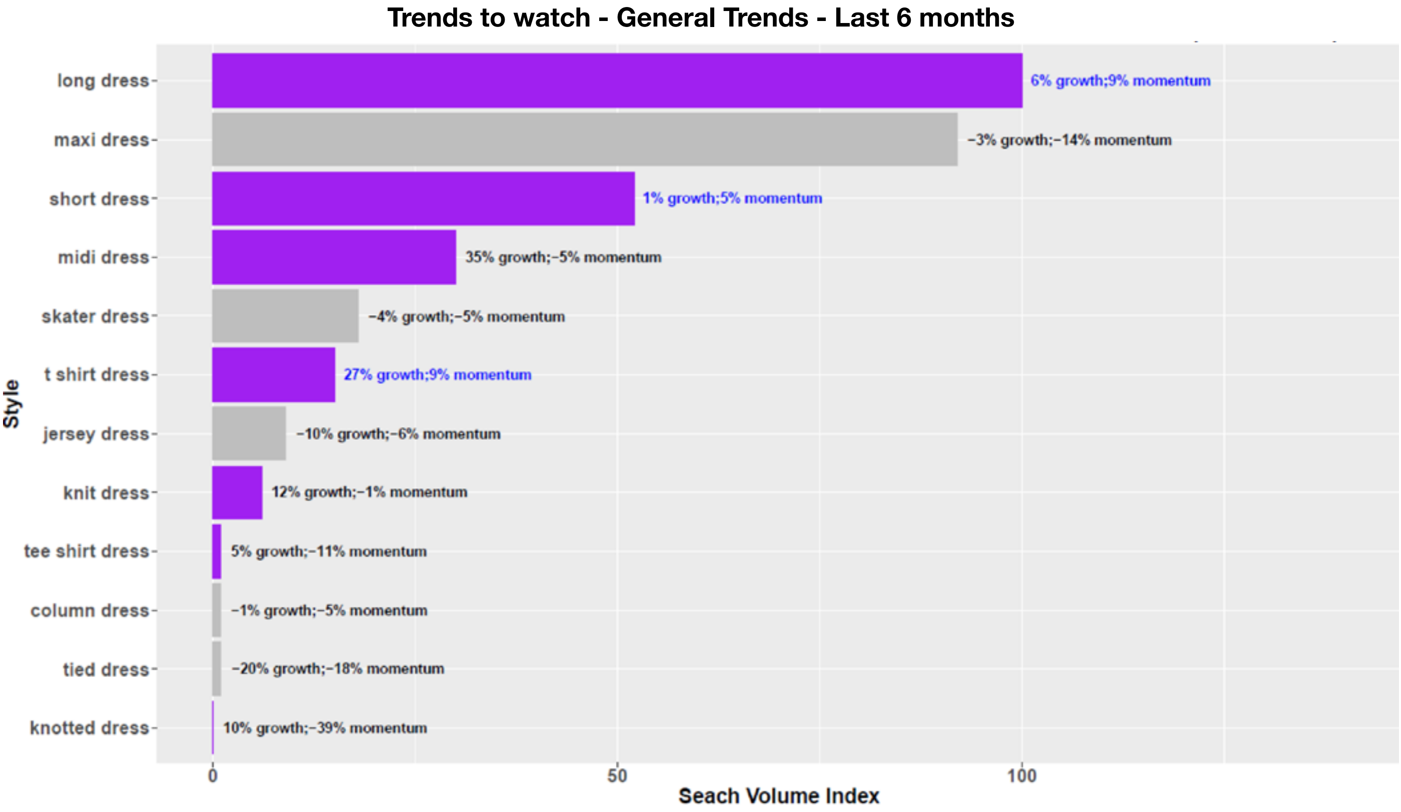


Figure 4: Rank of keywords for sampled styles

2.3. Cluster Keywords to Monitor Trends

Cluster analysis is a machine learning technique used to group keywords in such a way that objects within the same group (known as a cluster) are more similar to each other than to objects in other clusters. Clustering belongs to the category of unsupervised learning techniques, meaning that the objects we are dealing with are not explicitly labelled. The purpose of unsupervised learning is to discover hidden structures in data [4].

Figure 5 illustrates a dendrogram with agglomerative hierarchical clustering using complete link for 5-group clustering. The clustering algorithm chooses style pair whose merge has the smallest diameter. As this method takes the cluster structure into consideration, it is nonlocal in behaviour and generally obtains compact shaped clusters.

Based on the clustering results, **Figure 6** visualises the keywords using the popularity measuring metrics (volume), the long-term measuring metrics (growth) and short-term measuring metrics (momentum). In this two-dimensional scatterplot, x-axis indicates growth, y-axis indicates momentum, point size indicates volume, and each colour on the graph corresponds to a different style with a particular group.

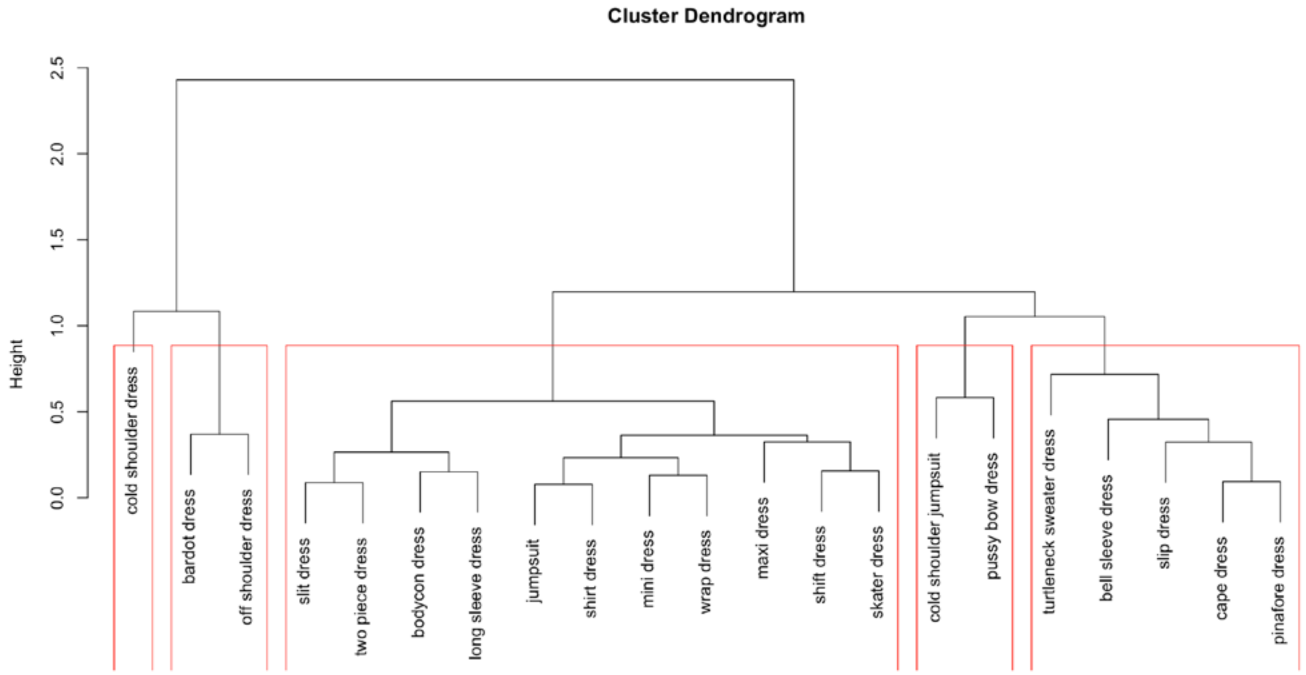


Figure 5: Hierarchical clustering results for sampled styles

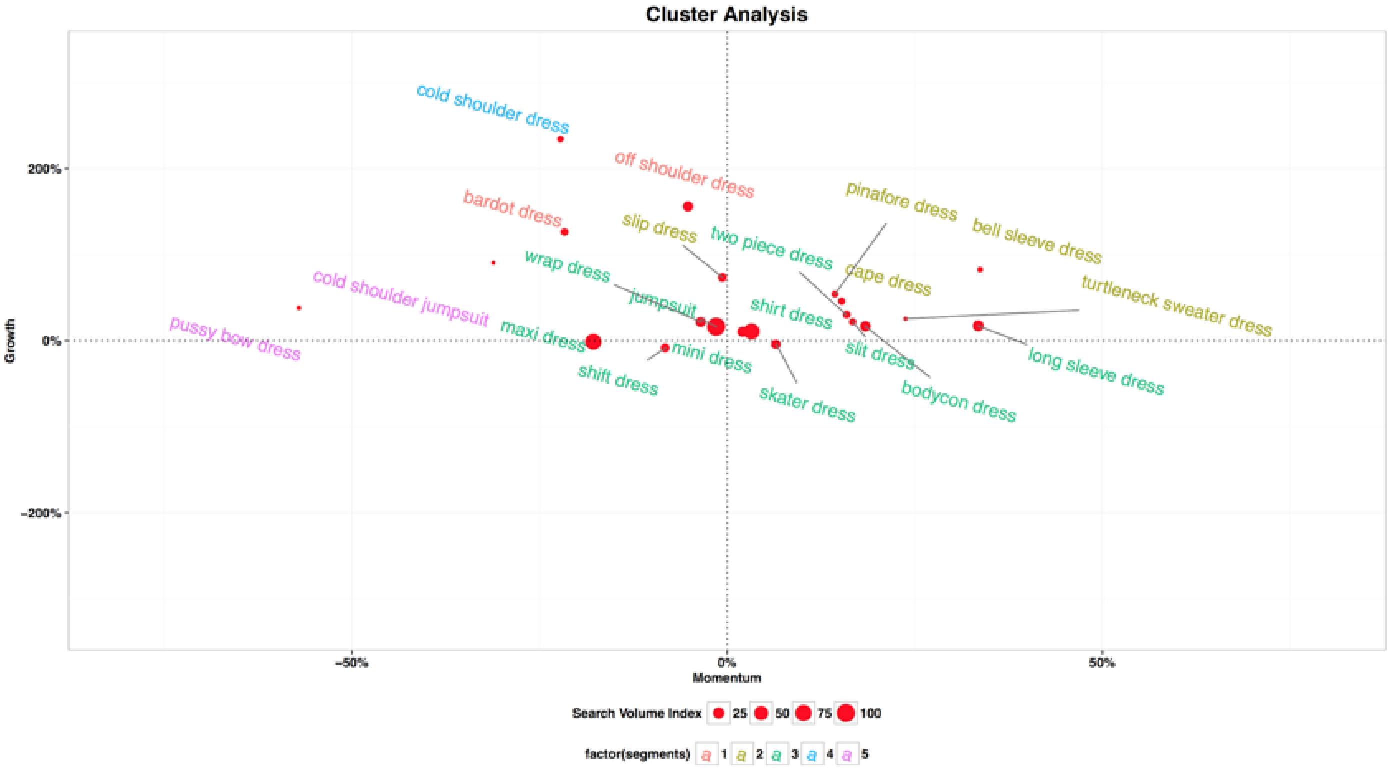


Figure 6: Visualisation of clustered results with volume, growth and momentum for sampled styles

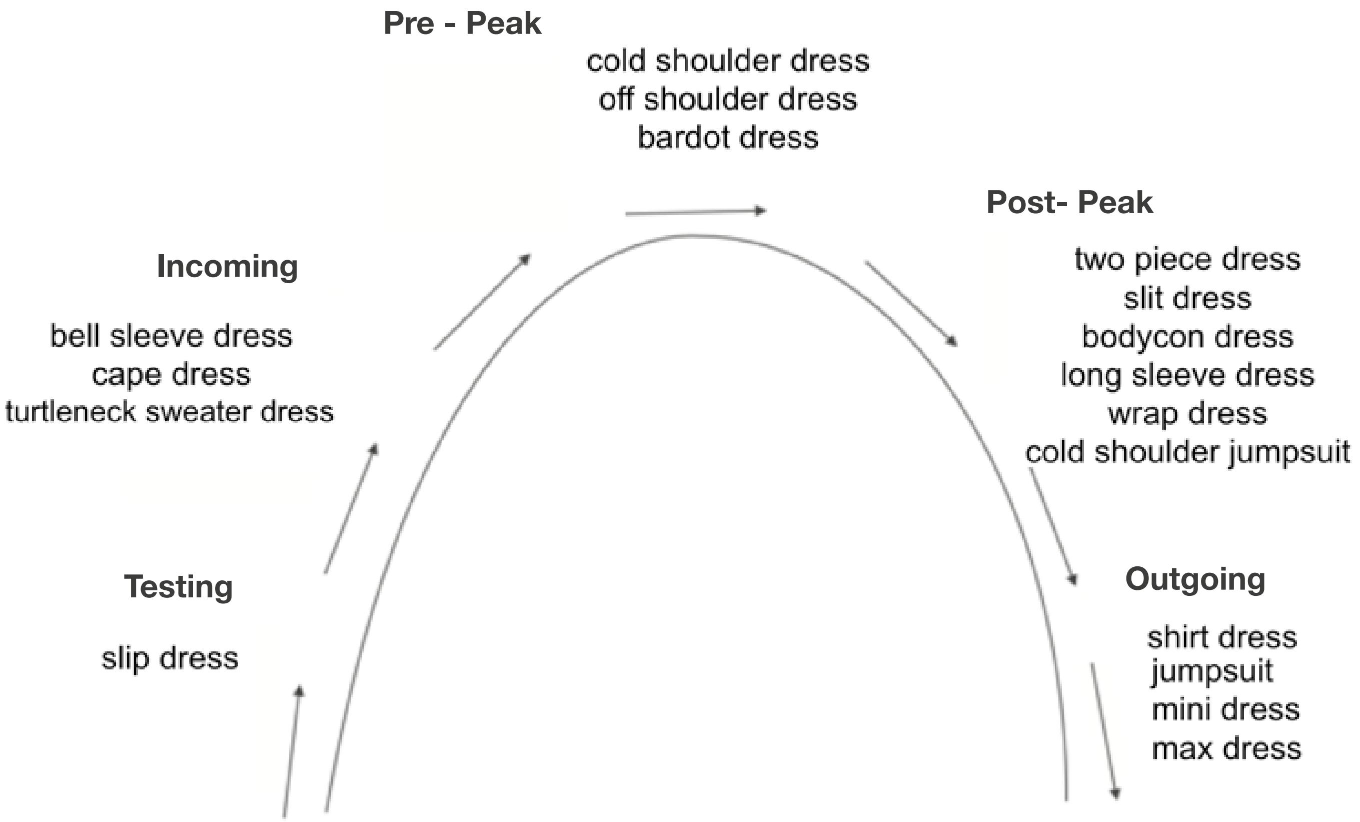


Figure 7: Future profiling clustered results for sampled styles based on business rules

A typical trending cycle includes ‘Testing’, ‘Incoming’, ‘Pre-Peak’, ‘Post-Peak’ and ‘Outgoing’ stages. Further profiling processes can break given styles into the above business groups. **Figure 7** illustrates the profiling results by using business rules. Take the following cases as examples:

- slip dress has 22% growth and 17% momentum but a small volume, which deserves to be watched closely as a ‘Testing’ style.
- bell sleeve dress ranks high, having 83% growth and relatively high momentum at 34%, making it an ‘Incoming’ trend.
- ‘Pre-peak’ stage includes cold shoulder dress having highest YoY growth but negative momentum.
- two-piece dress has the positive growth and momentum, naturally put into ‘Post-peak’ stage.
- shirt dress has -8% growth and -8% momentum, making it an ‘Outgoing’ trend.

Conclusions

More and more advertisers have a social media plan incorporated into their overall marketing plan. This paper introduces some applications to collect, aggregate, and mine social media data to generate actionable insights for product assortment, marketing and advertising. Given fashion styles, we can compile the top engaged posts based on social media buzz during study periods. Post data analysis is a fast way to learn what will shape the fashion business in the near future, featuring the top trend that popped and what they mean for business. Meanwhile, related keywords can be aggregated to calculate index data. The rank of trending keywords is important to show style popularity and imply customer interests. Predictive analysis on clustering styles can address user needs to offer data-informed, trend-driven products.

References

1. Barnes, N.G., and C. Daubitz. 2017. Time for Re-evaluation? Social Media and the 2016 Inc. 500. Center for Marketing Research, University of Massachusetts, http://www.umassd.edu/cmr/socialmedia-research/2017inc500/

2. Krystyanczuk, M. and Chatterjee, S..Python. Python Social Analytics. Packt Publishing. 2017.

3. Bonzanini, M.. Mastering Social Media Mining with Python. 2017.

4. Bishop ,C.. Pattern recognition and machine learning. Springer-Verlag, 2006.

Authors

Ling Huang is a Senior Research Scientist at Tumblr working with a Data Science and Analytics Team focusing on user interest and user behaviour analytics with big data. Ling holds her Ph.D in Statistics from Iowa State University. At Tumblr, she is working on many R&D initiatives connecting diverse social media data by leveraging the latest business trends and data science techniques including: machine learning, data mining and predictive models to create a holistic user experience.

ling@tumblr.com

Amanda Brennan is an internet librarian who specialises in researching tag data to better understand the inner workings of online communities. She has also done extensive work on meme culture, focusing on the historical trends within meme history and famous internet cats. In her current role as Tumblr's Senior Content Insights Manager, she runs The Fandometrics, which ranks fandoms in nine different entertainment categories. She holds a B.A. in English Literature from Drew University and a Masters of Library & Information Science, specialising in social media, from Rutgers University-New Brunswick. She blogs at memelibrarian.com and is found everywhere else as @continuants.

amandab@tumblr.com

Common Errors in Marketing Experiments and How to Avoid Them

Tanya Kolosova
Associates In Analytics Inc.

Samuel Berestizhevsky
Innovator and Actionable Analytics Expert

Classifications, Key Words:

- Design of experiments
- Marketing experiments
- Split-unit design
- Split-plot design
- Bias correction
- Block variables
- SAS

Abstract

A methodology for designing experiments developed by Sir Ronald Fisher is more than 80 years old, but many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. Because marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not only inefficient but are actually not suitable for conducting marketing experiments.

In this article, we describe the very useful and efficient split-unit (or split-plot) design of marketing experiments. Split-unit design is often used in marketing experiments but is not recognised; often missed or inappropriately analysed. This, in turn, produces misleading results that may be very costly in marketing. We use a real-life example to demonstrate some of the ideas involved and ways to correctly analyse split-unit design.

1. Introduction

A very common but inefficient approach to studying the effects of multiple factors is to carry out successive experiments in which the levels of each factor are changed one at a time (A/B testing). Sir Ronald Fisher showed that a better approach is to vary the factors simultaneously and to study response at each possible factor-level combination. A methodology of design of experiments (DoE) was developed by Fisher in his ground-breaking book “The Design of Experiments” in 1935. For his contribution in statistics, Fisher has been described as "a genius who almost single-handedly created the foundations for modern statistical science" (Hald, 1998) and "the single most important figure in 20th-century statistics" (Efron, 1998). Since then, DoE methodology has been broadly adopted in agricultural engineering, physical and social sciences, advertising and marketing.

Surprisingly, many marketers still rely on simple A/B tests to compare the performance of marketing campaigns and to find conditions to achieve the best results. There are multiple reasons to replace A/B tests by design of experiments:

a) In DoE, the approach is completely different from A/B testing, as all parameters (factors) are changed together, simultaneously, and not one parameter at a time. Thus, in DoE the required number of experiments is limited and significantly smaller than with A/B testing.

b) DoE provides a way to account for different sources of errors and compares averages to other averages rather than individual values to other individual values (as A/B testing). This achieves much greater accuracy in the estimation of effective factors for a given number of experiments, and thus the influential factors and their combinations are much more likely to emerge from the noise of the experimental errors.

c) But what is more critical, DoE allows for estimating of the impact of factor interactions which is not available in A/B testing. In fact, because marketing efficiency depends on a combination of factors and not on factors acting independently, A/B tests are not really suitable for conducting marketing experiments.

DoE methodology creates a framework for planning, analysing and executing marketing experiments. There are 3 main principles of DoE: randomisation, replication, and blocking. Randomisation is a deliberate process to eliminate potential biases from the conclusions through random assignment of “treatments”. Replication is, in some sense, the heart of all of statistics. Replication is the basic issue behind every method. We always want to estimate or control the uncertainty in our results. We achieve this estimate through replication; and blocking is a technique to include other factors in our experiment which contribute to undesirable variation. We want the unknown error variance at the end of the experiment to be as small as possible. Our goal is usually to find out something about treatment factors (or factors of primary interest), but in addition to this, we want to include any blocking factors that will explain variation.

One of the most efficient and frequently used designs is split-unit (also referred as split-plot) design: when one experimental unit is split into subunits, to which subsequent treatments are applied. Marketing usually involves a number of sequential steps, which makes split-unit design not only feasible and desirable but actually necessary.

The challenge is that split-unit experiments are often used but can be difficult to recognise. As a result, split-unit experiments are often inappropriately analysed. A spreadsheet of data can look like a variety of multifactor experiments, and it is very tempting to consider the experiment as completely randomised design (CRD) and then to apply straightforward analysis. In split-unit designs of experiments, it can take some research work to find out what factors (if any) are blocking factors and which are treatment factors, and (most importantly) what were the experimental units (EU) to which treatment factors were applied.

As with any statistical method, to receive correct results the method should be correctly applied. In the case of complex split-unit design, miss-interpretation of EU and incorrect error structure lead to inappropriate analysis, which produces misleading results that may be very costly in marketing.

Description of the Marketing Experiment

As an example, let’s consider the real-life case in which office supply retailing Company A needs to test the impact of marketing emails to find the optimal combination of factors-levels and achieve maximum sales as a response to marketing emails. To quantify the success of the marketing experiment, Company A uses total sales generated by the customers who participated in the marketing campaign.

There are multiple factors which affect the success of email marketing. For example, factors that describe the marketing message, format of the message, type of customers that receive the messages, etc.

In our real-life marketing experiment, the following 4 factors were included:

Factor Name	Levels	Factor Description
customer	C1 C2 C3	The Company A differentiates their customers into 3 types according to customers purchasing behaviour.
minimal_order	$50 $100	To become eligible for the discount, a customer has to make an order for a specific dollar value (at least).
discount	5% 10% 15%	If eligible according to the order dollar value, the customer receives a discount on the whole order.
subject_line	SL1 SL2	2 versions of email subject lines are developed by the marketers for the marketing experiment

First, lists of the 3 different types of customers were created. These lists were created by a random selection from the repository of the company's customers without replacement, which ensured that each selected customer appeared only once. Customers were selected according to customer types, producing 600,000 email recipients in each list. 4 replications of each type of customer were obtained, 12 Lists with 7.2million recipients in total.

Then, each List was randomly divided into 6 Batches of 100,000 recipients, and these Batches were randomly assigned combinations of the minimal order value and discount: ($50, 5%), ($50, 10%), ($50, 15%), ($100, 5%), ($100, 10%), ($100, 15%).

Next, each Batch was randomly divided into 2 Groups of 50,000 recipients each. Each Group was randomly assigned SL1 or SL2 version of email subject line.

The table below (the experimental table) presents the full factorial experiment 2^2 3^2 (36 treatment combinations) where each experiment cell contains 50,000 email recipients. The 4 replications of this experiment were conducted with an interval of 3 days.

Exp. run	customer	minimal_ order	discount	subject_ line
1	C1	$50	5%	SL1
2	C1	$50	5%	SL2
3	C1	$50	10%	SL1
4	C1	$50	10%	SL2
5	C1	$50	15%	SL1
6	C1	$50	15%	SL2
7	C1	$100	5%	SL1
8	C1	$100	5%	SL2
9	C1	$100	10%	SL1
10	C1	$100	10%	SL2
11	C1	$100	15%	SL1
12	C1	$100	15%	SL2
13	C2	$50	5%	SL1
14	C2	$50	5%	SL2
15	C2	$50	10%	SL1
16	C2	$50	10%	SL2
17	C2	$50	15%	SL1
18	C2	$50	15%	SL2
19	C2	$100	5%	SL1
20	C2	$100	5%	SL2
21	C2	$100	10%	SL1
22	C2	$100	10%	SL2
23	C2	$100	15%	SL1
24	C2	$100	15%	SL2
25	C3	$50	5%	SL1
26	C3	$50	5%	SL2
27	C3	$50	10%	SL1
28	C3	$50	10%	SL2
29	C3	$50	15%	SL1
30	C3	$50	15%	SL2
31	C3	$100	5%	SL1
32	C3	$100	5%	SL2
33	C3	$100	10%	SL1
34	C3	$100	10%	SL2
35	C3	$100	15%	SL1
36	C3	$100	15%	SL2

How Analysis Was Performed

This experiment was considered by Company A as a Completely Randomised Design (CRD) and analysed as such. The randomisation structure of the CRD implies that there is only one error term (the within error) and all factors effects are tested against it. The analysis was performed using a user-written computer program that utilises SAS® Software PROC MIXED (see SAS code with explanations in Appendix 1). The results are presented in the table below:

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	108	208.37	<.0001
minimal_order	1	108	0.57	0.4525
customer*minimal_order	2	108	2.08	0.1304
discount	2	108	10.65	<.0001
customer*discount	4	108	5.22	0.0007
minimal_order*discount	2	108	0.00	0.9956
customer*minimal_order*discount	4	108	1.29	0.2784
subject_line	1	108	1.61	0.2072
customer*subject_line	2	108	2.70	0.0717
minimal_order*subject_line	1	108	9.89	0.0021
customer*minimal_order*subject_line	2	108	0.69	0.5059
discount*subject_line	2	108	3.42	0.0364
customer*discount*subject_line	4	108	3.11	0.0183
minimal_order*discount*subject_line	2	108	2.53	0.0843
customer*minimal_order*discount*subject_line	4	108	2.17	0.0767

This table contains hypothesis tests for the significance of each of the fixed effects listed in the column "Effect". The following factors and their interactions were identified as significant (on 95% confidence level): customer, discount, customer*discount, minimal_order*subject_line, discount*subject_line, and customer*discount*subject_line.

Using significant factors, we built a regression model and found conditions (factors and their levels) that maximised response (sales). See SAS PROC MIXED code in Appendix 2. For each customer type (C1, C2, and C3), the conditions (factor-level combinations) that would generate maximum sales are presented in the table below:

customer	minimal_ order	discount	subject_ line	predicted sales
C1	$50	15%	SL1	$130,681
C2	$50	10%	SL1	$168,058
C3	$100	15%	SL2	$179,607

These results mean that if the email marketing campaign with the factors and levels presented in the above table is deployed for 50,000 customers of each type, then the Company A should expect, on average, the sales amount presented in "Predicted Sales" column.

How Analysis Should Be Performed

We suggest a closer look at how the experiment was executed to understand if the analysis of the experiment was performed correctly. First, customers were randomly selected by customer types, producing 12 Lists: 4 replications of each of 3 types of customers. This created a completely randomised design. Each List was an experimental unit (EU) for different types of customers (3 levels) – the entity to which types of customers are randomly assigned (see **Figure 1**).

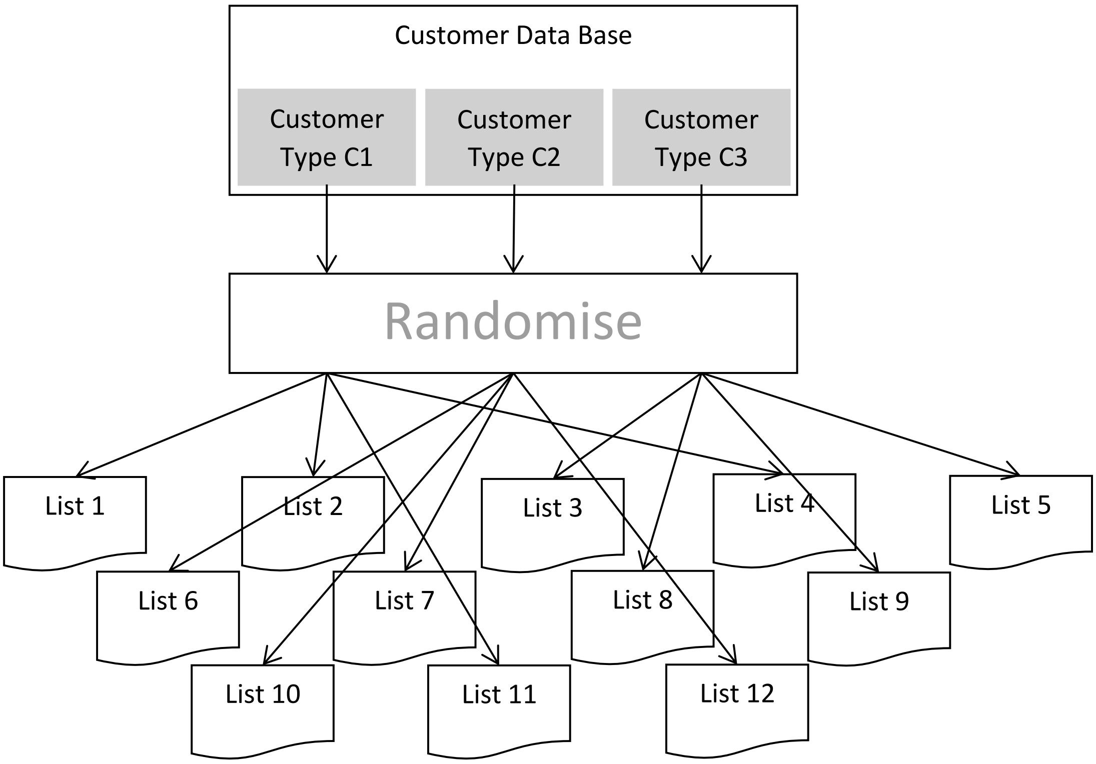


Figure 1. Lists Randomisation

Then, each List was randomly divided into 6 Batches. The act of grouping the experimental units together into homogenous groups is called blocking. Thus, the List was a block of 6 Batches, and the Batch was an experimental unit for combinations of the minimal_order and discount. In other words, the Batch design is a randomised complete block design, where the List is the blocking factor (see **Figure 2**).

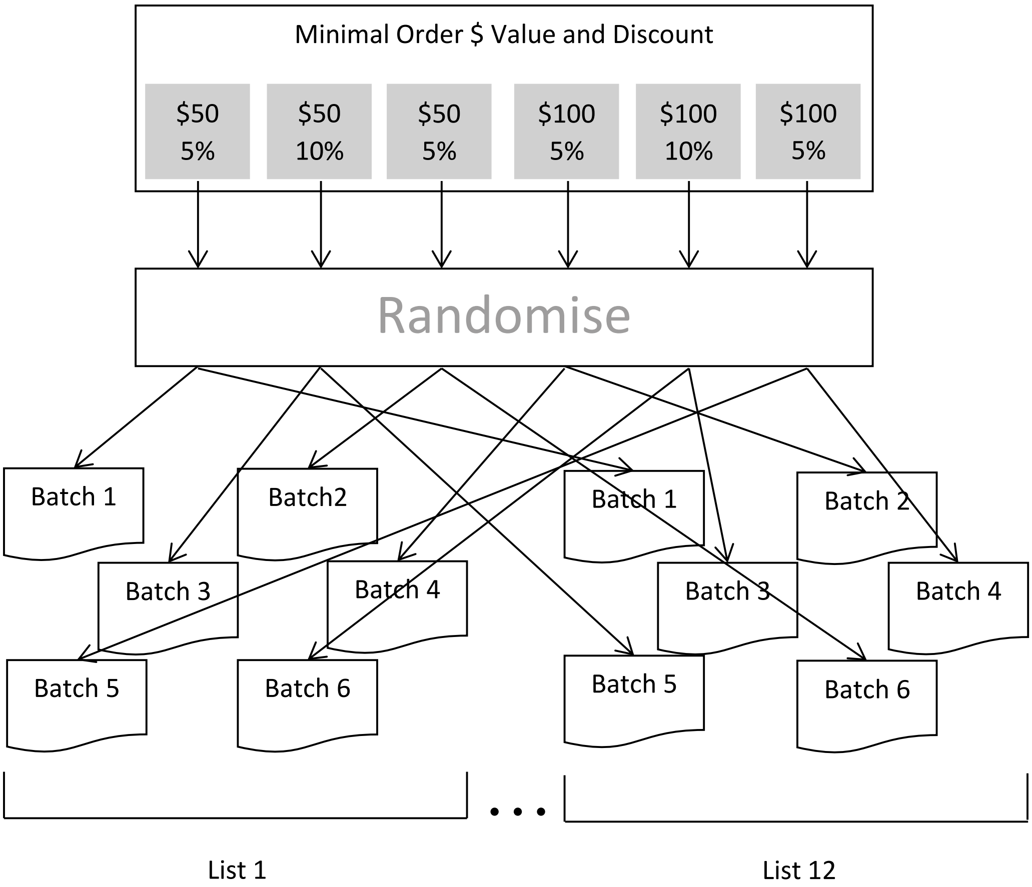


Figure 2. Batch Randomisation

And when each Batch was randomly divided into 2 Groups for 2 versions of email subject lines, Batch*List was a block for levels of email subject lines (see **Figure 3**).

Thus, the appropriate model includes:

- Factorial effects for levels of customer * minimal_order * discount * subject_line,
- and 3 experimental units: List, Batch, Group.

Using split-unit error structure, we analysed the results of the same experiment. SAS PROC MIXED code is presented in Appendix 3. Results of the analysis are presented below:

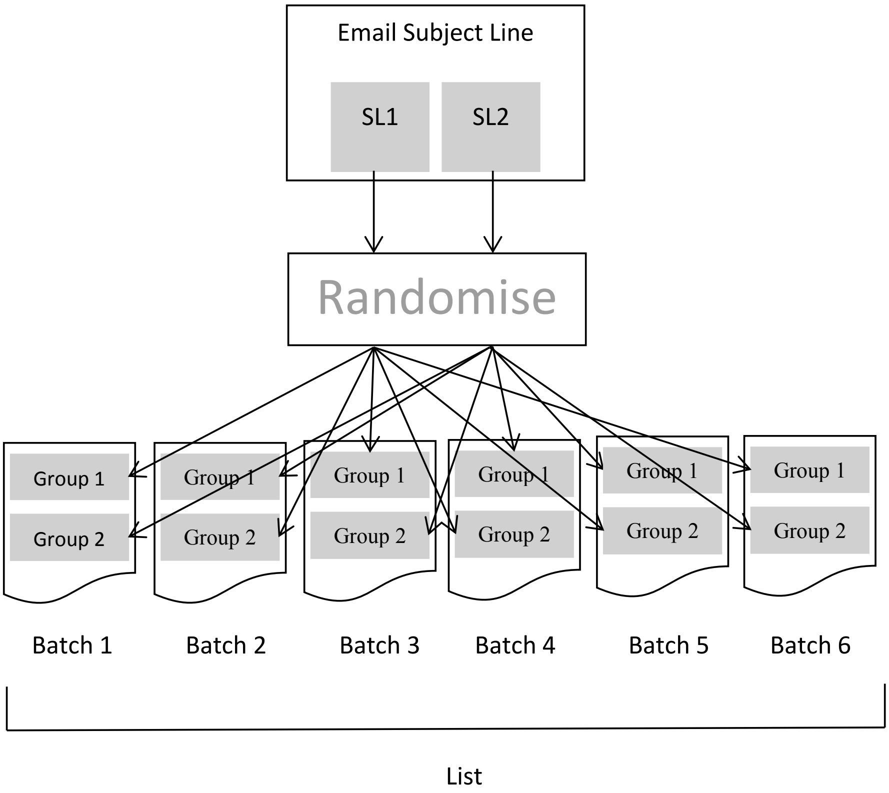


Figure 3. Groups Randomisation

Effect	Numerator DF	Denominator DF	F Stat	P-value
customer	2	108	208.37	<.0001
minimal_order	1	108	0.57	0.4525
customer*minimal_order	2	108	2.08	0.1304
discount	2	108	10.65	<.0001
customer*discount	4	108	5.22	0.0007
minimal_order*discount	2	108	0.00	0.9956
customer*minimal_order*discount	4	108	1.29	0.2784
subject_line	1	108	1.61	0.2072
customer*subject_line	2	108	2.70	0.0717
minimal_order*subject_line	1	108	9.89	0.0021
customer*minimal_order*subject_line	2	108	0.69	0.5059
discount*subject_line	2	108	3.42	0.0364
customer*discount*subject_line	4	108	3.11	0.0183
minimal_order*discount*subject_line	2	108	2.53	0.0843
customer*minimal_order*discount*subject_line	4	108	2.17	0.0767

The following significant factors and interactions were identified:

> customer, discount, subject_line, customer*discount, customer*subject_line, customer*discount*subject_line, minimal_order*discount*subject_line, and customer*minimal_order*discount*subject_line.

Now, we built a new regression model and estimated conditions (factor-level combinations) that maximised response (sales). See SAS PROC MIXED code in Appendix 4. For each customer type (C1, C2, C3), the conditions generating maximum sales are presented in the table below:

customer	minimal_ order	discount	subject_ line	predicted sales
C1	$50	10%	SL1	$140,174
C2	$100	10%	SL1	$156,830
C3	$100	10%	SL2	$191,097

In other words, if an email marketing campaign with factors and levels presented in the above table are deployed for 50,000 customers of each type, then Company A should expect, on average, the sales amount presented in "Predicted Sales" column.

Impact on the Business

The split-unit error structure allowed us to discover different interactions that existed in the experimental data. This is because the CRD analysis pools the three error terms – List, Batch, and Group – together, and the resulting error is not appropriate for any of the comparisons. In fact, the split-unit design is more complex, and it has more relationships among factors than CRD could discover.

CRD analysis found that the interactions minimal_order*subject_line and discount*subject_line were significant, while in reality, they were not. On the other hand, split-unit found that subject_line factor and interactions customer*subject_line, minimal_order*discount*subject_line and customer*minimal_order*discount*subject_line were significant, while CRD did not recognise it.

As a result, CRD analysis identified incorrectly the conditions (factor-level combinations) generating a maximum response (sales).

According to the CRD analysis, the best conditions for customer type C1 are 15% discount with minimum purchase of $50 while the email is sent with subject line SL1. These conditions should bring $130,681 in sales on average per 50,000 recipients. However, per our analysis, the model based on CRD analysis is incorrect. If we plug these conditions into the model that was built based on the split-unit analysis, the result will be $127,094, which is 2.7% less. If the campaign is sent to 1,000,000 recipients, it would translate to about $71,000 lower sales than expected.

For the same type of customers, the split-unit analysis identified conditions of 10% discount with minimum purchase of $50 while the email is sent with subject line SL1. Under these conditions, the expected sales from 50,000 of email recipients are $140,174. In comparison with the $127,094 that would be received under conditions identified by CRD analysis, the correct conditions would generate 10.3% more sales. And if the marketing emails with the conditions identified by split-unit analysis is sent to 1,000,000 recipients it would translate to $261,600 higher sales.

When we perform a similar examination for customer type C2, the results are the following:

- CRD analysis suggests that the best conditions (10%, $50, SL1) will generate $168,058.
- If we plug in these conditions into the predictive model based on the split-unit design, the

expected sales are $155,070, which is 7.73% less. Applied to a campaign for 1,000,000 recipients this will produce $259,760 less than expected.

- The split-unit analysis suggests that the best conditions (10%, $100, SL1) will generate $156,830. For 1,000,000 recipients this will produce $35,200 more than based on the conditions identified by CRD analysis.

For customer type C3, the results are the following:

- CRD analysis suggests that the best conditions (15%, $100, SL2) will generate $179,607.
- Plugged in into the split-unit model, these conditions will lead to $168,914 expected sales, 5.95% less. Applied to a campaign for 1,000,000 recipients this will produce $213,860 less than expected.
- The split-unit analysis suggests that the best conditions (10%, $100, SL2) will generate $191,097. For 1,000,000 recipients this will produce $443,660 more than expected from CRD conditions.

Summary

Design of Experiment applied to marketing helps identify factors and their interactions that maximise a marketing campaign's performance (sales or customer purchases).

Failure to identify the appropriate design structure leads to an incorrect analysis of the experiment, and as a result, produces misleading inferences.

Using the real-life example, we demonstrated how to analyse a marketing experiment and identify correct error structure. We showed how to incorporate the split-unit error structure, perform appropriate analyses and build correct predictive models. The comparison of results obtained from CRD vs. split-unit design demonstrated immediate impact on business performance.

References

1. Box, G.E.P., Hunter, W.G., Hunter, J.S. (1978). Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. New York: Wiley.

2. Efron, B. (1998). R. A. Fisher in the 21st century. Statistical Science, 13: 95–122.

3. Hald, A. (1988). A History of Mathematical Statistics. New York: Wiley.

4. Kenward, M., Roger, J. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics 53, 983-997.

5. Kolosova, T., Berestizhevsky, S. (1998). Programming Techniques for Object-Based Statistical Analysis with SAS Software. Cary, NC: SAS Institute Inc.

6. Littell, R.C.L., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

7. Wu, C.F.J, Hamada, M. (2000). Experiments: Planning, Analysis, and Parameter Design Optimisation. New York: Wiley.

Tanya Kolosova is an expert in the area of actionable analytics and analytical software development having served as the Senior Vice President of Research and Analytics at IPG Inc, Principal Researcher at Yahoo!, Vice President of Analytics at Nielsen and Chief Analytics Officer at YieldWise Inc. Tanya developed her expertise with extensive depth and breadth of experience in bringing mathematical disciplines to bear on marketing and other business problems. She has extensive knowledge of audience intelligence, design and analysis of marketing experiments, market-mix modelling, and multi-channel commerce and has worked in a variety of industries like online and offline retail, telecom, finance, and more. Tanya also co-authored two books on statistical analysis and metadata-based applications development with SAS, which are used in universities globally and she was featured in Forbes Magazine (2006) for her work for GAP. In 2017 Tanya co-founded InProfix Inc, a stealth mode startup that develops AI solutions for the insurance industry. She is currently a Principal at Associates In Analytics Inc.

Samuel Berestizhevsky is an innovator and actionable analytics expert having served as the Chief Technology Officer at YieldWise Inc. Samuel has extensive knowledge of software development methods and technologies; artificial intelligence methods and algorithms; as well as statistically designed experiments. He has developed and deployed analytical software solutions for a variety of industries like online and offline retail, telecom, finance, and more. Samuel co-authored two books on statistical analysis and metadata-based applications development with SAS which are used in universities globally and was featured in Forbes Magazine (2006) for his work for GAP. In 2017 Samuel and Tanya Kolosova co-founded InProfix Inc, a stealth mode startup that develops AI solutions for the insurance industry.

Appendix 1

The following statements of PROC MIXED fit the completely randomised design model.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer|minimal_
order|discount|subject_line;
run;
```

The dataset experiment contains the experimental table described in the article. The variables replication, customer, minimal_order, discount, and subject_line are listed as classification variables in the CLASS statement.

Customer|minimal_order|discount|subject_line listed on the right side of the MODEL statement mean that the model is built of all possible combinations of these factors. The dependent variable sales is listed on the left side of the MODEL statement.

Appendix 2

The following statements fit the completely randomised design model and estimate prediction according to this model.

```
proc mixed data=experiment cl;
class replication customer min_order discount
subject_line;
model sales=customer discount customer*discount
customer*subject_line
      minimal_order*subject_line
customer*discount*subject_line
      /solution singular=1e-11 ddfm=kr
outpm=prediction;
run;
```

Variables and their combinations listed on the right side of the MODEL statement contain all effects that were identified as significant at the previous step. ddfm=kr means that the degrees-of-freedom method of Kenward and Roger (1997) is in effect. outpm=prediction requests to create the dataset with predicted sales values.

Appendix 3

The following statements fit the split-plot model assuming random block effects.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer|minimal_
order|discount|subject_line;
random replication(customer) minimal_
order*discount*replication(customer);
run;
```

Variables and their combinations listed in the RANDOM statement define random block effects.

Appendix 4

The following statements fit the split-plot model with random block effects and estimate prediction according to this model.

```
proc mixed data=experiment cl;
class replication customer minimal_order
discount subject_line;
model sales=customer discount subject_line
customer*discount customer*subject_line
      customer*discount*subject_line minimal_
order*discount*subject_line
      customer*minimal_order*discount*subject_
line
   /solution singular=1e-11 ddfm=kr
outpm=prediction_split;
random replication(customer) minimal_
order*discount*replication(customer);
run;
```

`outpm=prediction_split` requests to create the dataset with predicted sales values according to split-plot model.

Who is Who with Behavioural Data

AN ALGORITHM TO ATTRIBUTE THE DEVICE'S NAVIGATION TO USERS SHARING THE SAME DEVICE

Carlos Ochoa
Netquest

Carlos Bort
xplore.ai

Josep Miquel Porcar
Netquest

Classifications, Key Words:

- Passive data
- Behavioural data
- Browsing information
- User identification

Abstract

Passive data is powerful but still faces many challenges to gain trust as a way to understand people's online behaviours. One major challenge is separating the data from several individuals sharing one single browsing device. Existing solutions to overcome this difficulty are clearly unsatisfactory. A new method to separate navigation data without asking users, preserving the passive nature of the data, is explored in this paper.

Introduction

Data fuels market research; insight generation is impossible without relevant and appropriate data to support it. Fortunately, we are seeing tremendous growth in the amount of new sources and the diversity of data available for research, at a low cost that used to be unimaginable.

Any technological disruption is as much an opportunity as it is a challenge, no matter the field we are talking about. The market research industry is experiencing a major disruption, and it is no exception to this rule.

For many years, the number of ways of accessing consumer data was limited and stable. With the advent of the internet, researchers started to adopt this new channel to access potential respondents. It was at the beginning of this century and, since then, things have rapidly evolved. In developed markets, most of the data is collected by means of online access panels (ESOMAR, 2016; Baker et al., 2013; Lozar-Manfreda & Vehovar, 2008).

However, despite this considerable progress in the way consumers are accessed, the nature of the collected data has not evolved at the same pace. Offline surveys are increasingly replaced by online surveys; traditional focus groups sometimes are replaced by online focus groups; and the same may be said of many other traditional methods. In other words, the internet has made data collection more cost efficient and fast, but not radically different. At least, up to now.

After nearly 15 years, it is only now that we are witnessing a real revolution in the way researchers use the internet, motivated by several factors: widespread adoption of social media, irruption of the mobile internet and consolidation of e-commerce, among others. And, of course, the learning curve of the internet adoption has moved ahead. As a result, new data types are now available and new methodologies are being developed on top of them.

Passive online data collection has emerged as one of the most promising, groundbreaking methodologies. One of its most powerful variants is the installation of an online meter on the browsing devices of members of an online access panel to record information on their online behaviours (visited websites, apps usage, search terms), as well as their opinions (via survey).

Passive data has proved to have an edge over survey data when researching online behaviours, overcoming (1) human memory limitations and (2) lack of sincerity (Revilla, Ochoa, Loewe and Voorend, 2015). However, passive online data collection still faces several challenges that prevent broader adoption. First, the large amount of data generated per individual makes the analysis complex; additionally, some of the uses being given to such data require new analytical methods, as the traditional are facing significant constraints. Second, individual navigation may be spread across different devices (smartphone, tablet, personal PC, professional PC, etc.) which would require installing a meter on all the individual's devices to get the full picture. Finally, some browsing devices are shared among several users, preventing us knowing for certain which browsing information comes from each one.

This latter issue produces serious discomfort to researchers, as it is an objective distortion of the data. Existing solutions do not enjoy widespread support due to several drawbacks: reduced representativeness, increased measurement error or lack of transparency on how they work. In fact, some solutions may be worse than the problem they try to solve.

This paper presents a completely new approach to overcome the user identification problem: separating individuals' navigation by means of an algorithm that just looks at the data. As it will be shown, succeeding in doing so is only possible if browsing information is a personal trait, something unique that unequivocally identifies each individual the same way a fingerprint does (PII).

This paper is organised into several sections:

In **section I**, information on the data used to carry out this research is shared.

In **section II**, existing solutions and their limitations are reviewed.

In **section III**, the key hypothesis that must be valid to make our purpose possible is detailed (i.e. the way each individual browses the internet is unique), as well as some data supporting it.

Section IV provides a detailed description of the proposed algorithm, while section V shares its results.

In **section VI** we will explore how some limitations of this solution could be overcome, suggesting further research to improve results.

Section I. The Data

Data Source

We use data from the Netquest's Behavioural Panel in US, UK and Spain. The final algorithm was trained on data from the Spanish panel, as it has been recording behavioural data for a longer period of time.

These behavioural panels are built on top of existing online access panels, so online behavioural and survey data can be collected from the same sample of individuals. To do so, a subsample of the access panel is invited to install tracking software (from now on called the "meter") on their browsing devices (PCs, tablets and smartphones). The meter collects data on the individuals' online activity, such as URLs of the visited webpages, time of the visits, and app use in the case of mobile devices.

As all the metered panellists are also members of the survey panel, their basic sociodemographic information is known as well as some profiling data on different topics (e.g. automotive, healthcare, FMCG, etc.). When regular panellists are invited to install the meter, they are asked to complete an installation survey that asks how many devices they use to browse the internet and, for each device, (1) type of device, (2) main use of the device (personal/professional) and (3) whether it is shared or not. Panellists can install the meter on all their devices and they are rewarded for each one (up to three different devices). However, they are not obliged to track all their devices.

We are interested in panellists that have installed the meter on a shared PC or tablet. Although mobile devices can be shared occasionally, they are mainly single-person devices.

Definitions

Behavioural data produced by a meter is a record of visited webpages, like the one shown in **Figure I**.

Start data and time	Webpage URL
2016-03-04 T19:04:48	http://www.google.com
2016-03-04 T19:04:56	https://www.google.com/search?q=bestsellers+2017
2016-03-04 T19:05:25	https://www.amazon.com/
2016-03-04 T19:05:42	https://www.amazon.com/gp/site-directory
2016-03-04 T19:05:58	https://www.amazon.com/books-used-books-textbooks/b/
...	

Figure I. An example of the behavioural data collected by a meter, also known as clickstream. Data has been simplified: additional metadata is also recorded, such as device type, user id, etc.

For the sake of clarity, we define here two key words that will be used throughout this paper.

Webpage/URL: A webpage is a particular file on the internet that can be accessed by an individual through a browser. A webpage is described by a URL, an address that univocally identifies a webpage (www.amazon.com/help/display.html). The terms webpage and URL will be used interchangeably from now on.

Website/Domain: A website is a connected group of webpages regarded as a single entity, under the same domain name. So, https://www.amazon.com/gp/site-directory and https://www.amazon.com/books-used-books-textbooks/b/, are two webpages under the same website, described by the domain name amazon.com. The terms website and domain will be used interchangeably from now on.

The Dataset

We have data available from our target group (shared PCs and tablets). However, we cannot produce a validation dataset to train and test an algorithm. For our purpose, a validation dataset would be a collection of visited webpages from a shared device, each webpage properly labelled as belonging to the right user. Without a validation dataset, it is not possible to measure how accurately an algorithm separates navigations.

This is precisely one of the main obstacles of this work: the lack of a validation dataset. In order to get one, we should rely on some of the already existing methods to separate navigations; however, those methods' accuracy is under suspicion.

To overcome this difficulty, an artificial validation dataset was used. It was built by mixing two individuals' navigations from non-shared devices, as if both individuals were sharing the same device. Knowing who is the real author of each webpage visit allows us to use this dataset to measure how well a classification algorithm performs at identifying the right user.

In particular, the artificial dataset used to test the algorithm was built by a two-stage sampling process:

- Stage 1: a random sample of N=200 individuals was drawn from the Spanish Behavioural Panel.
- Stage 2: 1,000 couples of navigations were mixed by randomly selecting couples of panellists from the initial sample.

This approach limits the validity of the results that have been obtained:

- Two real users sharing the same device might have navigations more similar than two random users selected from the panel. In other words, this work aims to contribute to solving the shared devices issue by proving that two independent user's navigations can be separated; the next step will be to test this solution on two navigations coming from the same device.
- We have focused our research on separating two navigations, while a browsing device might be shared by three or more individuals.

In the final section, some considerations will be shared on how these limitations could be overcome.

Sessions

The algorithm described in section IV uses the concept of browsing session. A session is defined as a group of successive webpage visits, in a way that the time lapse between consecutive visits is shorter than a timeout parameter.

As will be detailed later, the classification algorithm assumes that all the visits within a session belong to the same user. This information is key for the accuracy of the output. So, the timeout (time difference to consider a new session) is a tuning parameter of the model and not an intrinsic feature of the data; that is, we need to define the timeout in the most convenient way to maximise the accuracy of the algorithm.

Tuning parameters such as the timeout can be adjusted using different solutions: "A general approach that can be applied to almost any model is to define a set of candidate values, generate reliable estimates of model utility across the candidates' values, then choose the optimal settings." (Kuhn and Johnson, 2010).

The parameter tuning process should be part of the algorithm creation. The optimal timeout value: (1) places as many webpage visits in the same session as possible, but (2) limits the risk of grouping together visits from different users. This could be called the information-precision trade-off.

However, we cannot find the optimal timeout through our artificial data. Our dataset is made up from pairs of independent navigations mixed together, so it does not provide relevant information to find the right balance in the information-precision trade-off.

In view of that fact, we decided to use a timeout of 30 minutes, following a standard used by popular analytical tools such as Google Analytics (https://support.google.com/analytics/answer/2731565). The resulting sessions were manually inspected, ensuring that the division of the webpage visits among sessions made sense.

Section II. Existing Solutions

To our knowledge, three main solutions have been proposed to overcome the shared device issue.

The **first solution is to limit the data collection to non-shared devices**. This way, misclassification is avoided but at the expense of reducing the data availability and introducing sample coverage error (i.e. people sharing devices may be different from people not sharing devices).

The **second option is to add a "login dialog" to the meter**; so, each time the user starts a browsing session (or the browser has been inactive for some time), a pop-up message asks about his/her identity. Theoretically, this solution ensures that each webpage visit recorded by the meter is attached to the right user. In practice, serious doubts arise on the reliability of this identification method. Ultimately, asking people about their identity while they use the internet may violate the passive nature of the data, producing: increased churn rate of the participants and misreported identities due to both lack of attention when using the login dialog and social desirability. One of the goals of collecting data passively is to observe people's activity without affecting their behaviours; by adding a login dialog this benefit might vanish.

Finally, some companies claim that they identify the user behind the device **by analysing his/her keyboard keystroke pattern**. These companies do not disclose details on how this technique works and how well it performs, a fact that may cause distrust among researchers. Even if we accept this technique is truly effective, the rapid evolution of browsing devices may challenge further development: new touch keyboards, autocomplete features in the address bar of the browsers, voice typing, etc. On top of that, browsers are evolving towards greater control of which data is shared with third party applications; so, in the future, using metadata (such as keystroke data) might be problematic.

Technique	Pros	Cons
Avoid non-shared devices	• Simplicity. • No misclassification errors	• It skips the problem, does not solve it. • Representativeness issue: lack of data from shared devices. • Valuable data is not used.
Login dialog	• It would be perfect… if users were perfect. • Easily applicable to more than two users.	• It violates the passive nature of the data: people are constantly aware of being tracked. People may hide part of their navigation. • Users' misuse (lack of attention when selecting the identity in the login dialog) may produce more harm than good.
Analysis of the keyboard keystroke patterns	• Non-intrusive.	• Lack of transparency on how it works. • Challenged by devices' interface evolution.

Table I. Pros and cons of existing solutions to separate navigations.

Table I summarises the pros and cons of each solution. A new approach, like the one proposed in this paper, would enjoy clear advantages over the existing solutions:

- Simplicity: separation is achieved by just inspecting the data.
- Pure passive: panellists are not asked to provide information while navigating.
- Robustness against future technological limitations that may restrict the possibility of collecting metadata.

Lastly, a final thought about how these solutions might evolve in the future: new technical capabilities can make separation techniques unnecessary. Passive facial recognition is a good candidate; Apple is the most recent technology company who is utilising facial recognition to identify device users to unlock phones. Such systems could be used to separate navigations. However, it may be perfectly possible as well that such information is not available for third party applications running on the device, as is currently happening with other sensitive information.

Section III. Main Hypothesis

To what extent is the way you browse the internet different from the way others do? This is the key question whose answer determines whether our approach makes sense or not. And, of course, our initial hypothesis is that the answer to this question is yes.

A simple exploration of the data at hand (sample from the Behavioural Netquest Panel described in section I) helps to support this hypothesis (**Table 2**).

People in the sample visited 175.2 different domains on average in a month, ranging from 4 to 915. But the relevant fact for our purpose is that a pair of randomly chosen panellists from the sample only share 4% of their unique visited domains.

This fact, that supports our hypothesis, could seem somewhat surprising. It is well known that websites such as Google and Facebook capture most of the Internet traffic in the Western world. In fact, 98.5% of the panellists in the sample have visited Google while 86.0% have visited Facebook in the time period under analysis. However, that does not mean that people only browse such popular websites. **Figure 2** shows which percentage of the panellists have visited (at least once) each of the different domains that are present in the data.

Figure 2 follows a typical Pareto distribution: just a few domains are visited by most of the panellists, while there is a long tail of less popular domains visited by a small part of the sample. Popular domains play a minor role to distinguish users; rare domains are the ones that can be crucial to achieving our goal. Our algorithm aims to exploit this opportunity.

Sample size N	200
Country	Spain
Type of device	Non-shared PC or Tablet
Time period	1 month (June 2016)
For the whole sample…	
Sessions	77,842
Visited webpages	1,259,076
Visited domains	363,929
Unique visited domains	17,309
For each panellist average [minimum – maximum]	
Sessions	389.2 [5 ↔ 1,345]
Webpages per session	16.2 [1 ↔ 1,708]
Domains per session	4.7 [1 ↔ 67]
Unique visited domains in one month	175.2 [4 ↔ 915]
For each couple of panellists…	
% shared domains	4.0% [0% ↔ 25%]

Table 2. Dataset description. Data for panellists and couples of panellists is shown in the format “average [minimum ↔ maximum]”.

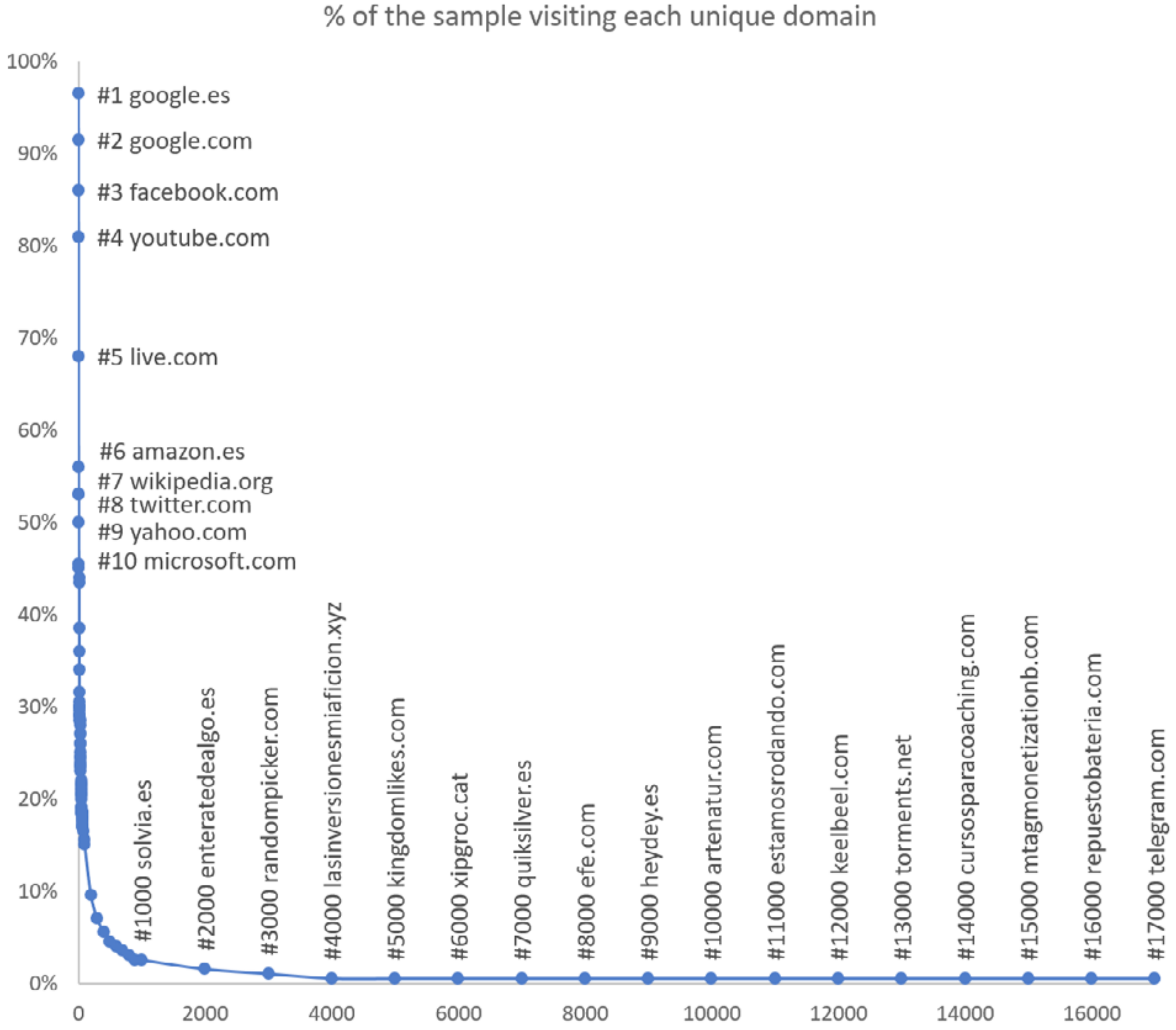


Figure 2. Popularity of each website among the sample. Just a few websites are visited by most of the panellists while a long tail of domains are visited by very few panellists.

Section IV. The Algorithm

Choosing learning type

Our purpose is to design a classification algorithm that gets as input the mixed navigation from two users (A and B) and returns each visited webpage properly labelled as belonging to user A or B.

A first decision to be made in pursuing this goal is deciding what type of algorithm should be implemented. Machine Learning literature usually classifies algorithms in three broad categories, depending on the way the algorithms learn: supervised, unsupervised and reinforcement learning algorithms.

Reinforcement learning algorithms were rapidly discarded. Reinforcement is a powerful learning method, but such algorithms are "not given examples of optimal outputs (…) but must instead discover them by a process of trial and error" (Bishop, 2006). This learning process requires a reward system: the algorithm needs to know if each step made contributes or not to a success metric. But unfortunately, our problem does not provide such rewards.

On the other hand, supervised learning requires training data that "comprises examples of the input vectors along with their corresponding target vectors" (Bishop, 2006). This is precisely what we lack; and what we have tried to replace with the artificial data described in section I.

Supervised learning assumes that a causality relationship exists between some input factors and the classification criteria. Such learning offers some advantages compared to unsupervised learning. There are many and powerful supervised algorithms at hand that have proved to perform extremely well in a wide variety of problems (Kuhn and Johnson, 2013): Classifications Trees, Random Forest, Boosting, Support Vector Machines, etc.

However, after a few attempts to train one of these algorithms, it soon became evident that it was not the right approach. It was pretty easy to train a supervised algorithm to separate the navigation from two particular individuals by using a specific training dataset from these individuals. But the result cannot be generalised to other pairs of individuals, a problem known as overfitting (Kuhn and Johnson, 2013). In other words, if we had data from two specific individuals correctly classified, we could train a specific model for them to classify their future navigation. But this model cannot be used for other individuals.

One key learning from this unsuccessful attempt is the following: while each individual uses the internet in a unique way, it is not easy at all to predict which is this unique way based on personal characteristics.

So, in light of the evidence, an unsupervised algorithm was the only option available. As explained by Bishop (2006), "In other pattern recognition problems, the training data consists of a set of input vectors x without any corresponding target values. The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering (…)". That description fits perfectly with our problem: regardless of the factors that explain why people browse some websites rather than others, we just want to identify groups of "similar websites" in the hope that this will reveal the identity of the individuals behind.

The solution: an ensemble of different algorithms

A description of the unsupervised algorithm developed to separate navigations is provided below, step by step. All the data manipulations and algorithms have been programmed in R (www.r-project.org), using public libraries.

To facilitate the reader's comprehension, the description is backed with real examples.

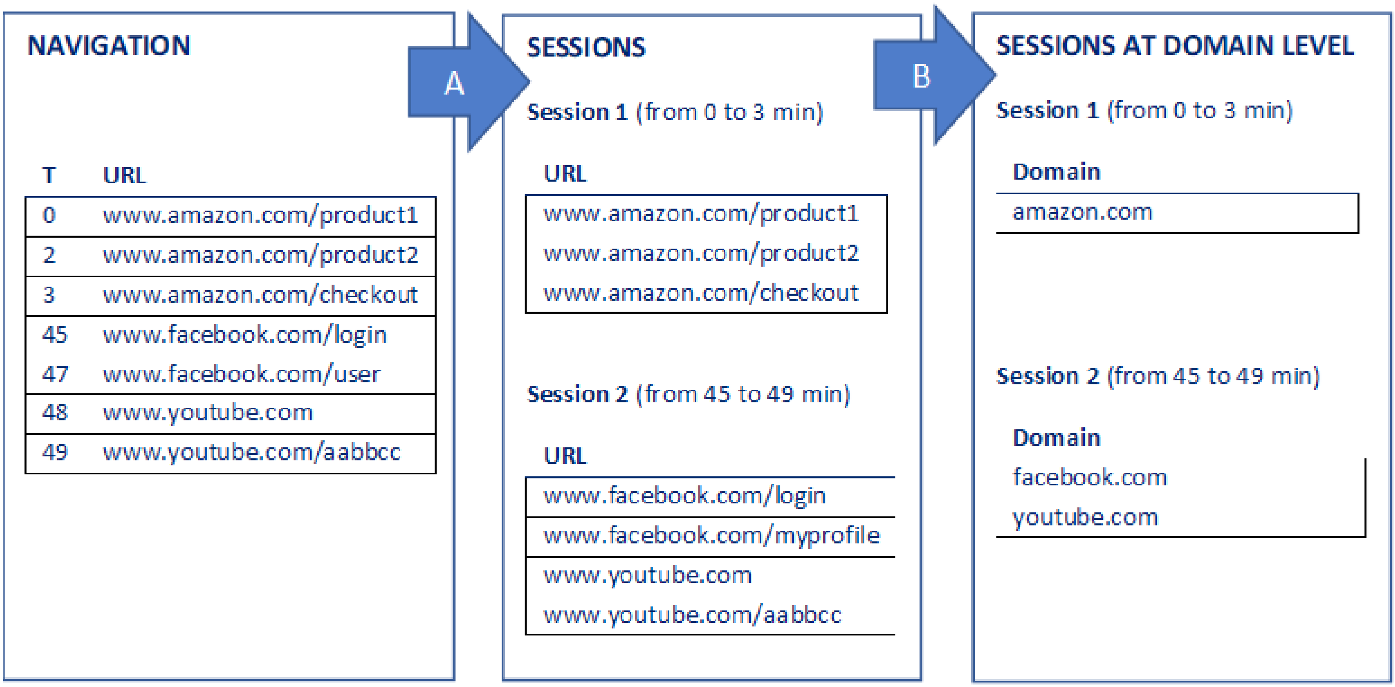
Step 1: Dimension reduction

Navigation data consists of a list of complete URLs, each one formed by a domain name (e.g. amazon.com), sometimes a subdomain (e.g. aws.amazon.com, www.amazon.com) and a page descriptor (e.g. www.amazon.com/cell-phones-service-plans-accessories/b/ref=nav_shopall_wi).

We have decided to focus our analysis on domain names. For our purpose, all this information around the domain name (i.e. subdomains and page descriptors) adds much more complexity. As the whole idea behind the separation is to exploit coincidences in the same session, working with precise URLs would require much more data to train an algorithm.

The same applies to multiple visits per domain. We could potentially separate two users visiting the same domain if one user tends to visit many pages in the same session and the other one just a few. But this information is much less relevant than the simple fact of whether a domain has been visited or not, so we decided to not use it.

So finally, data is reduced to domain level: a complete navigation is transformed in a list of sessions (**Figure 3 – A**), and each session is transformed into a list of unique domains visited in that session (**Figure 3 – B**). Once the session is properly assigned to the right user, all the suppressed information around the domain can be recovered in order to assign webpage visits to each user.



NAVIGATION

T	URL
0	www.amazon.com/product1
2	www.amazon.com/product2
3	www.amazon.com/checkout
45	www.facebook.com/login
47	www.facebook.com/user
48	www.youtube.com
49	www.youtube.com/aabbcc

SESSIONS

Session 1 (from 0 to 3 min)

URL

www.amazon.com/product1
www.amazon.com/product2
www.amazon.com/checkout

Session 2 (from 45 to 49 min)

URL

www.facebook.com/login
www.facebook.com/myprofile
www.youtube.com
www.youtube.com/aabbcc

SESSIONS AT DOMAIN LEVEL

Session 1 (from 0 to 3 min)

Domain

amazon.com

Session 2 (from 45 to 49 min)

Domain

facebook.com
youtube.com

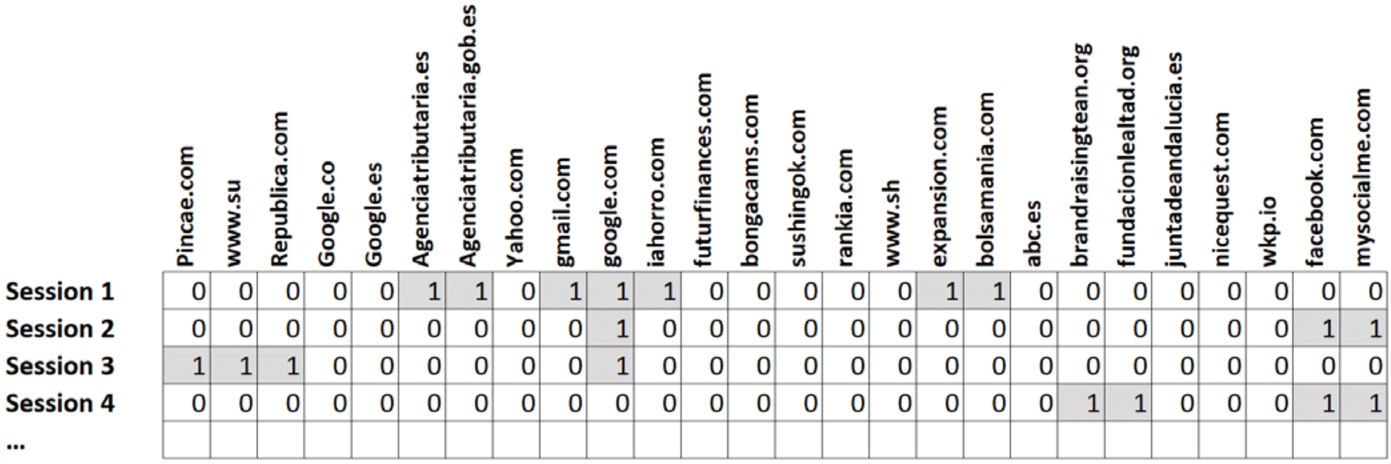
Figure 3. Dimension reduction: navigation data is first grouped into sessions; then, URLs are reduced to unique domains per session.

Step 2: Similarity matrix

A similarity matrix is evaluated for the list of unique domains present in the navigation data. Each cell of this matrix contains a measure of how likely two domains appear together in a session, which is a sort of correlation.

Several ways to create such matrix were tested, without relevant differences in the result. So, the following simple method was finally employed:

1. First, the browsing data for each couple of panellists is binary coded in a matrix M. This matrix has as many rows as sessions and as many columns as unique domains in the joint navigation. So, each row is a sequence of ones and zeros that represents whether each possible domain is present at each session (**Figure 4**).



	Pincae.com	www.su	Republica.com	Google.co	Google.es	Agenciatributaria.es	Agenciatributaria.gob.es	Yahoo.com	gmail.com	google.com	iahorro.com	futurfinances.com	bongacams.com	sushingok.com	rankia.com	www.sh	expansion.com	bolsamania.com	abc.es	brandraisingtean.org	fundacionlealtad.org	juntadeandalucia.es	nicequest.com	wkp.io	facebook.com	mysocialme.com
Session 1	0	0	0	0	0	1	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Session 2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Session 3	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Session 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
...																										

Figure 4. Matrix M binary codifies sessions and unique domains.

2. Once ***M*** is built, the similarity matrix ***S*** is calculated by means of a matrix product ***M***^{T}***M***. The similarity matrix ***S*** has the following properties:

a. Each cell represents the similarity of a pair of domains. So, the matrix is symmetric.

b. The minimum value of each cell is zero, that means that both domains never appear together in a browsing session (minimum similarity).

c. The maximum value of each cell is the number of times both domains appear together in a browsing session.

d. The diagonal of that matrix represents the similarity of each domain with itself. Because of the matrix ***M*** is computed, each diagonal cell equals the number of times each domain appears in the navigation. It can be removed as it does not provide useful information.

Domains that appear together more frequently in the sessions will score high in the similarity matrix ***S***. As we assume that sessions are owned by a single user, high similarity indicates high likelihood of belonging to the same user. A visualisation of such a matrix is shown in **Figure 5** for a couple of individuals.

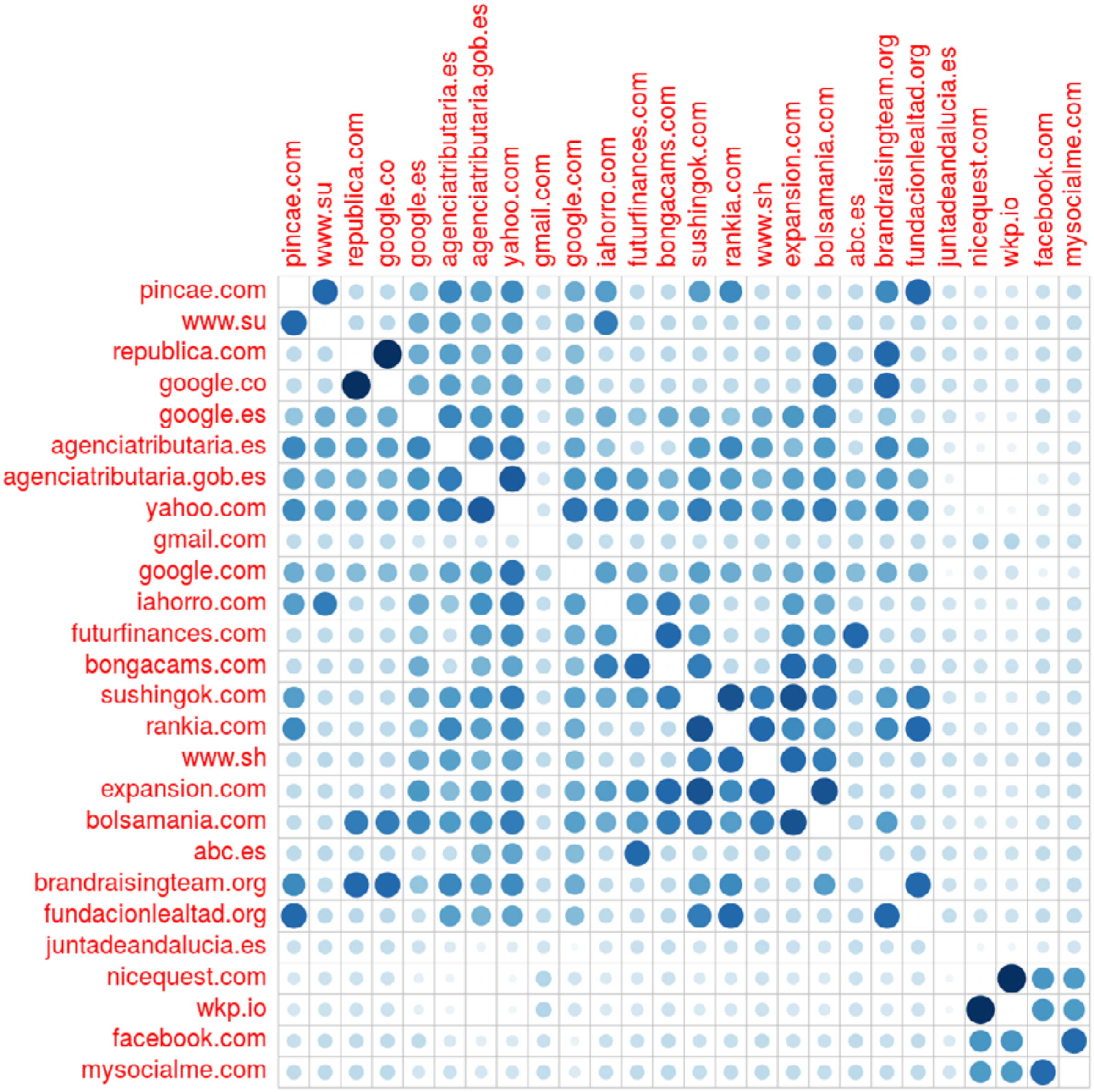


Figure 5. Visual representation of a similarity matrix of unique domains for a couple of individuals. Dark areas mean high similarity between domains; that is, domains that frequently appear together in the same sessions.

Theoretically, the above procedure to create the similarity matrix has some drawbacks. For instance, domains that appear more often tend to score higher. In other words, the matrix is not normalised. For instance, say that a pair of domains A and B appear just two times in the navigation but always together, while another couple of domains C and D appear tens of times but only two times together. Both pairs A-B and C-D will get the same similarity score (2), while it seems clear that A-B are more similar than C-D.

Different strategies to overcome this alleged limitation were tested. Even though some of these strategies produced similarity matrices in greater accordance with what we may expect, none of them improved the final performance in the ultimate goal of the algorithm: correct separation of individual’s navigations.

Step 3. Multidimensional Scaling

The similarity matrix ***S*** tells which domains appear together more often and which ones do not. But we aim to combine this pair wise information to get a global picture. Could we spread the domains onto a plane, so the similar domains are placed together and the dissimilar ones are placed distant? If so, we could separate two groups of domains in the hope that each group belongs to a different individual.

An intuition on how domains can be placed in a plane in such a way is the following:

1. First, the similarity matrix ***S*** can be transformed into a distance matrix ***D*** by inverting each cell one by one (***D***=1/***S***). So instead of a measure of similarity, we get a measure of dissimilarity: the higher the value in a cell, the less likely two domains appear together in a session, the less likely they both belong to the same individual.

2. Place the first unique domain in the centre of a plane.

3. Take the second domain and place it at a distance from the first one according to the information in the distance matrix D.

4. Things get more interesting with the next domains; if you try to proceed in a similar way as with the second domain, the distance with the first and the second domain must be considered at the same time. But both distances may be incompatible, so a compromise among different distances must be reached.

5. This process gets increasingly complex as more domains are placed in the plane because the coherence between pair wise distances gets harder. So, for instance, when placing the 6th domain, its distance with the 1st, 2nd, 3rd, 4th and 5th domains must be balanced.

Fortunately, this placement of domains in a physical space can be done using a well-stablished algorithm, a Multidimensional Scaling (MDS; Young, 1987). An MDS algorithm aims to place several elements in a N-dimensional space such that the between-elements distances are preserved as well as possible. Each element is then assigned coordinates in each of the N dimensions.

How a MDS works can be seen with a simple example. Consider the distances between nine American cities (**Table 3**).

Running a MDS on this data, a pair of coordinates is assigned to each city. Once represented in a graph, the position assigned to each city approximately reproduces the shape of the US map (**Figure 6**). In other words, the best representation on a 2-dimensional space of the distances between cities is the United States of America.

The same MDS procedure can be applied to the between-domains distance matrix ***D***. As we want to identify just two groups of domains, we can use a 1-dimensional plane (N=1), that is, a simple line. If the hypothesis supporting this work is valid, domains that are visited only by user A should be placed at one end of the line, while domains visited only by user B should be placed at the opposite end of the line. Domains that are visited by both users should be placed at the half-way.

Considering the midpoint of the line (coordinate x=0) as the threshold to separate both users, the distance to this midpoint can be considered a propensity score. Domains with large scores (positive or negative) are highly likely to be visited by only one of two users; these are highly **discriminant domains**. On the contrary, domains with scores close to zero are likely to be shared between both users and therefore, **non-discriminant**.

	BOS	CHI	DC	DEN	LA	MIA	NY	SEA	SF
BOS	0	963	429	1,949	2,979	1,504	206	2,976	3,095
CHI	963	0	671	996	2,054	1,329	802	2,013	2,142
DC	429	671	0	1,616	2,631	1,075	233	2,684	2,799
DEN	1,949	996	1,616	0	1,059	2,037	1,771	1,307	1,235
LA	2,979	2,054	2,631	1,059	0	2,687	2,786	1,131	379
MIA	1,504	1,329	1,075	2,037	2,687	0	1,308	3,273	3,053
NY	206	802	233	1,771	2,786	1,308	0	2,815	2,934
SEA	2,976	2,013	2,684	1,307	1,131	3,273	2,815	0	808
SF	3,095	2,142	2,799	1,235	379	3,053	2,934	808	0

Table 3. Distance between nine US cities

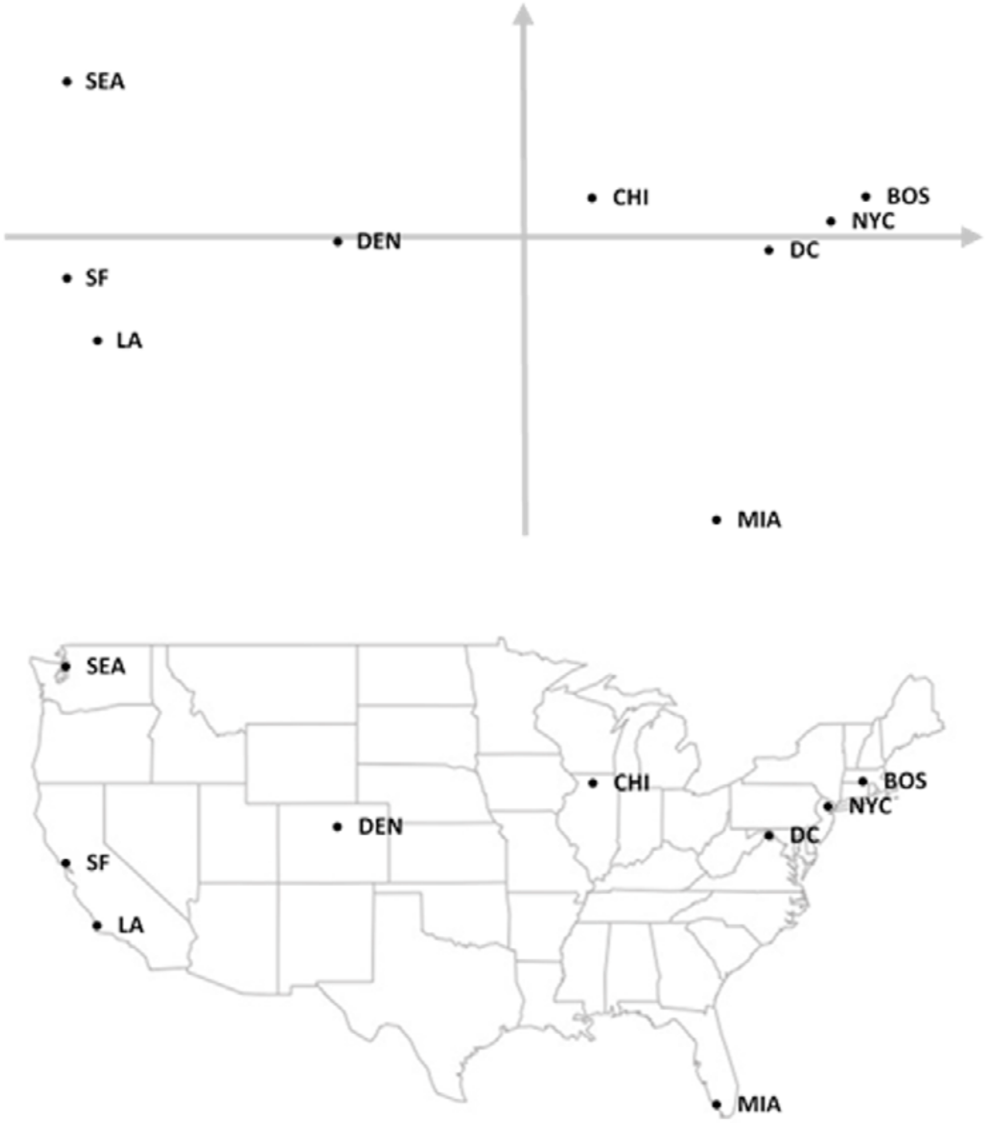


Figure 6. The MDS procedure locates the cities in a plane in a way that between-cities distance is preserved as much as possible. This results in the real relative location of each city.

Figure 7 shows an example of the resultant scores for a couple of users.

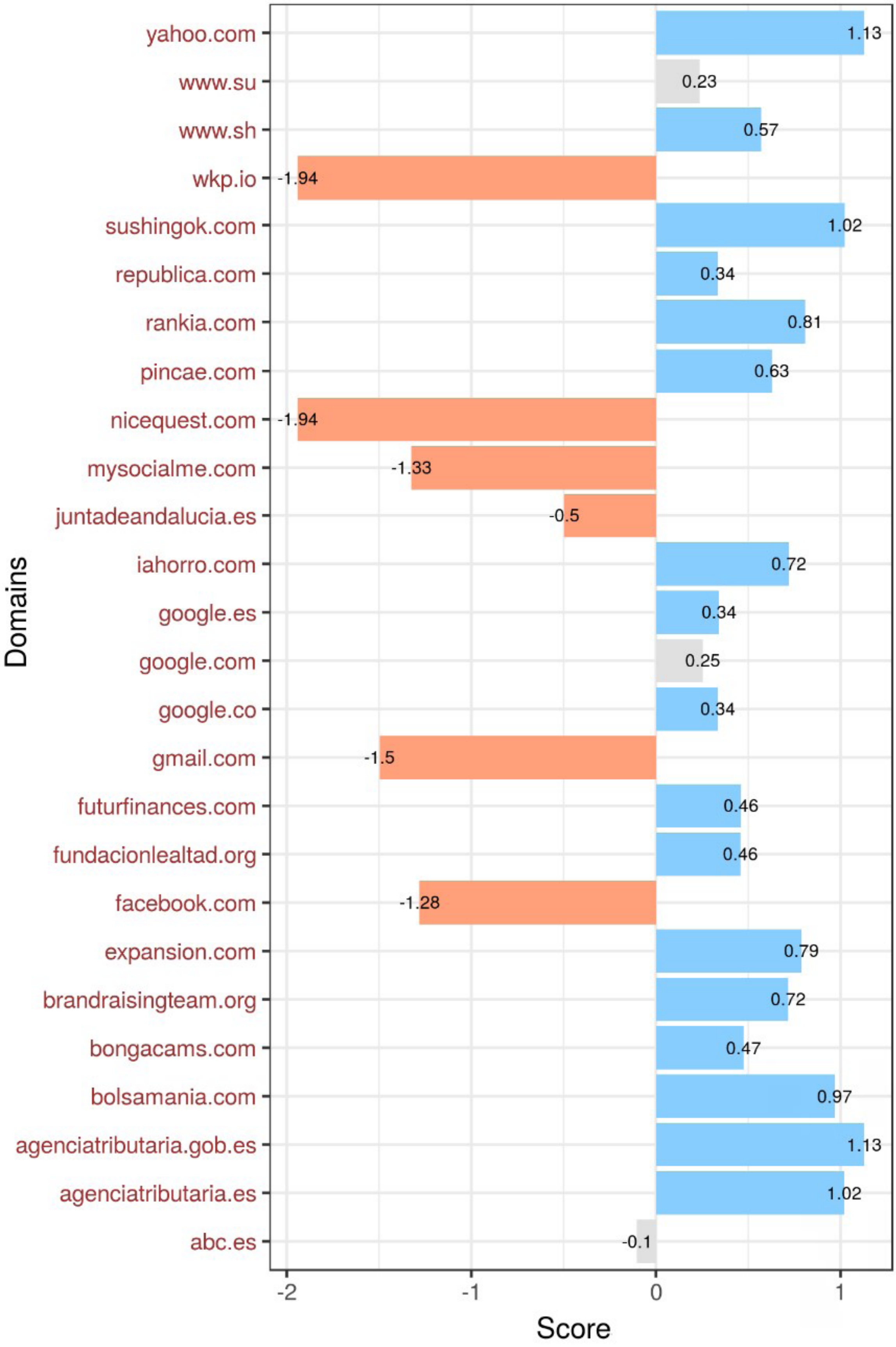


Figure 7. Propensity scores of a pair of users. Large scores are associated to discriminant domains (blue and orange), while small scores (grey) to non-discriminant ones.

Step 4. Classification of sessions

Up to step 3 we have found a first classification criterion at domain level: positive scores are assigned to user A and negative ones to user B. However, this criterion is too extreme: even close-to-zero score domains (non-discriminant) are assigned the same way that large score domains are. For instance, according to the data shown in **Figure 7**, the domain google.com (score +0.25) should be assigned always to user A, while it is likely that both users visit google.com.

The accuracy can be improved by computing average scores per session. For instance, consider a session with four domains (facebook.com, gmail.com, mysocialme.com and republica.com) with respective scores -1.28, -1.5, -1.33 and 0.34. Consider also that positive scores are assigned to user A and negative ones to user B. When assigned at domain level, facebook.com, gmail.com and mysocialme.com are assigned to B and republica.com to A. But if we evaluate the average score of the session (-0.9), the four domains are assigned to B (see **Figure 8**).

This procedure allows us to assign non-discriminant domains more accurately, taking advantage of the fact that they are part of a session that may include discriminant domains. The larger a session is, the more effective is this method. Of course, short sessions with non-discriminant domains are more likely to be misclassified. Fortunately, short sessions impact much less in the global accuracy.

Step 5. Who is the panellist?

One final step is missing. Up to step number 4 we have separated domains in two groups, A (positive scores) and B (negative scores). But, who is the user we are interested in?

If we certainly know that the user we are interested in (target user) visits a particular discriminant domain, this is enough to decide. We call this domain, the one that is visited exclusively by the target user, the hook. If the **hook** is positive, the target user is A, otherwise is B.

But, how can we get such a domain from the target user? Fortunately, when installing the meter on an online panel such as Netquest, a hook is always available: the domain of the panel website, the one accessed when the panellist participates in surveys. As such a domain is only visited by the target user, is the perfect hook.

And what about if both users sharing a browsing device are members of the panel? Theoretically, both users would visit the panel domain and it would no longer be a hook. However, we can benefit from the fact that panellists need to log in the panel website to participate in surveys.

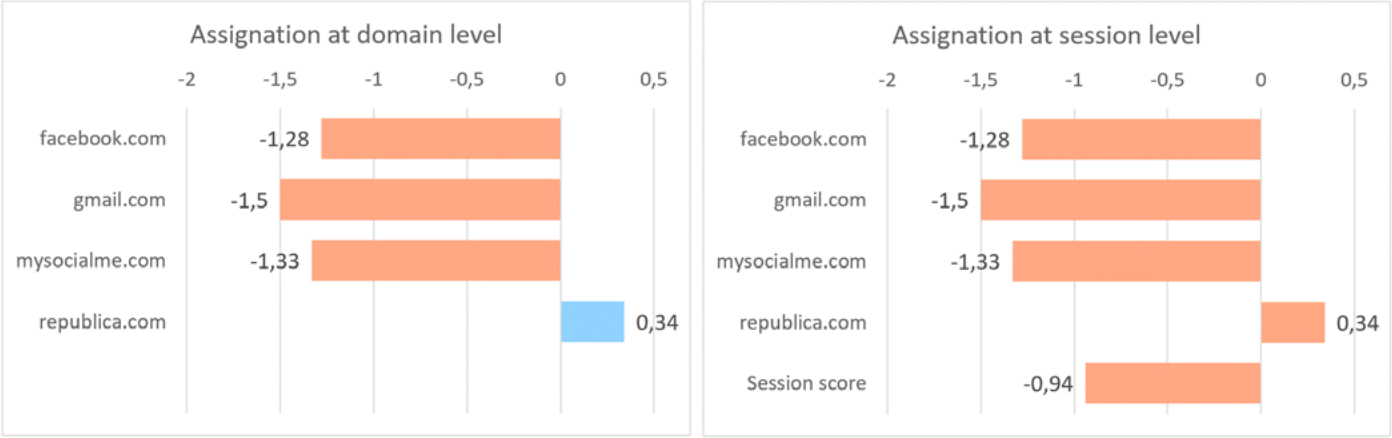


Figure 8. When considered at domain level, each domain is assigned based on its score, even those with small scores. When considered at session level, all four domains are assigned together. In this case, republica.com is assigned to user B (orange) despite its positive score.

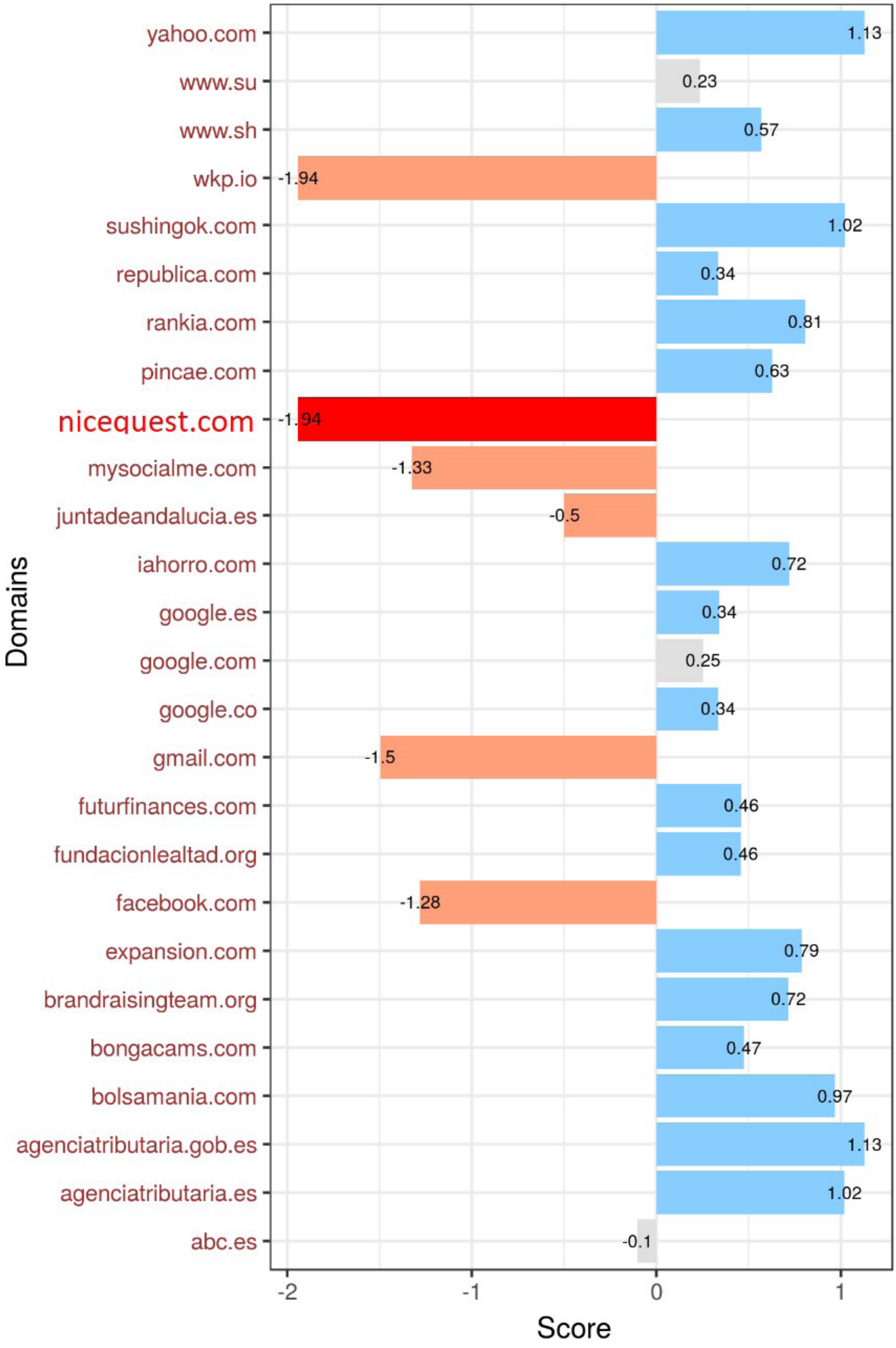


Figure 9. The sign of the hook's score determines whether the target user is A or B. In this example, the hook (nicequest.com) determines that the target user is B (the one with negative scores).

When users log in into a website, the URL changes specifically for each user, so we can still use the panel domain as a hook; or more precisely, two slightly different versions of the panel domain. For instance, the domain nicequest.com becomes "nicequest.com/userA" for user A, and "nicequest.com/userB" for user B. So, in fact, when both users are panellists, we have two different hooks at our disposal; which makes user identification even more reliable.

Section V. Performance

In order to assess how well this algorithm performs, a success metric must be defined. Several options are available:

- Session accuracy: the percentage of sessions (as defined in section I) assigned to the right user.

- Domain accuracy: the percentage of unique domains per session assigned to the right user. This metric can be applied directly to the data once the dimension reduction is applied (see section IV, step 2).

- URL accuracy: Each domain in the dataset corresponds to several URLs (different webpages from the same domain that the user visits in the same session). Once the domains are classified, the underlying URLs are implicitly classified, resulting in the URL accuracy metric. URL accuracy may differ from domain accuracy only if (1) some domains receive more page visits than others and (2) these domains have a significant different accuracy.

We have computed all the above metrics, but we have used the domain accuracy as the key success metric, the one used to compare algorithms' performance.

Table 4 shows the different resultant metrics for our dataset.

Success metric	Average
Session accuracy	84.0%
Domain accuracy	87.3%
URL accuracy	87.5%

Table 4. Performance of the algorithm (different metrics)

A naïve classifier (that is, assigning domains randomly to each user) may reach a 50% domain accuracy, so this is the base accuracy we aim to improve. Considering this fact, 87.3% domain accuracy should be considered a promising result.

Note that the domain accuracy is greater than the session accuracy (87.3% vs 84.0%). This is due to the fact (explained in section IV, step 4) that larger sessions tend to be better classified. URL accuracy, on the contrary, is pretty similar to domain accuracy, meaning that domains that receive more webpage visits are not better classified.

Figure 10 shows the accuracy for each pair of users in the dataset, ordered by accuracy. It is interesting to note that the average accuracy (87.3%) is not the result of a homogenous performance among all the cases; on the contrary, the algorithm seems to work very well (accuracy > 85%) for a near 75% of the cases, while dramatically fails in some cases. In these cases, the algorithm performs even worse than a naïve random classification algorithm.

Looking at these problematic cases some insights are revealed:

- In some cases, the hook has been mistakenly classified. The result of this misclassification is harmful: the accuracy goes towards zero (in fact, towards 1 minus the accuracy that should be achieved if the hook were properly classified). A hook misclassification may occur when (1) the panellist has visited the panel website very few times (e.g. he/she has participated in few surveys) and (2) those visits have occurred in very short sessions.

- Some particular cases were found in which the pair of mixed navigations seemed to be produced by three different users instead of two. This fact can be easily visualised by executing the MDS algorithm using two dimensions (a plane) instead of one (a line). **Figure 11** shows one of these cases. When doing so, domains are displayed in three different areas that can be easily separated. A possible explanation is that although we are producing the artificial dataset by combining navigations from devices that are allegedly not shared, some of them are actually shared. Some users may have reported misleading information in the installation survey regarding the shared condition of their devices, or this condition has changed after some time. In fact, this finding reinforces the approach we have taken: relying more on the data than on declared information.

Section VI. Further Research and Applicability

Our research has been tested on artificial data, as detailed in section I. Despite results that are promising, an objection could be made on the fact that we have not tested the algorithm on navigations coming from a real shared device. If people sharing a device navigate in a much more similar way, the main hypothesis that supports this work would be seriously compromised.

A real validation dataset from shared devices would be needed. But we should not underestimate how difficult to obtain one is: whenever we ask

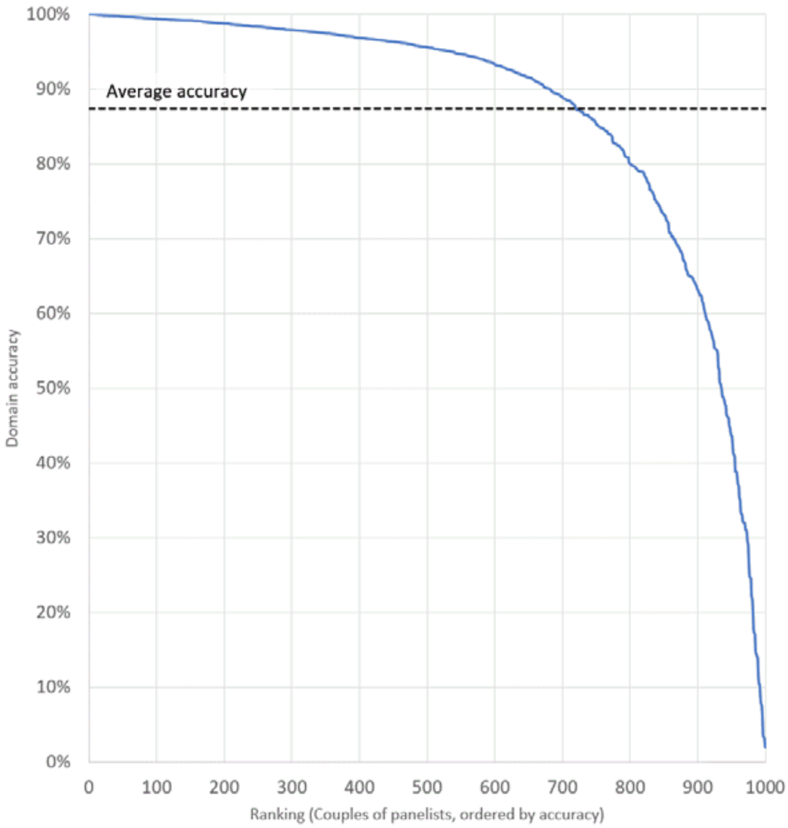


Figure 10. The overall accuracy is a mix of many well classified cases and a few very bad classified ones. If we order the 1,000 cases, most of them perform between 90% and 100% but some perform lower than 50%.

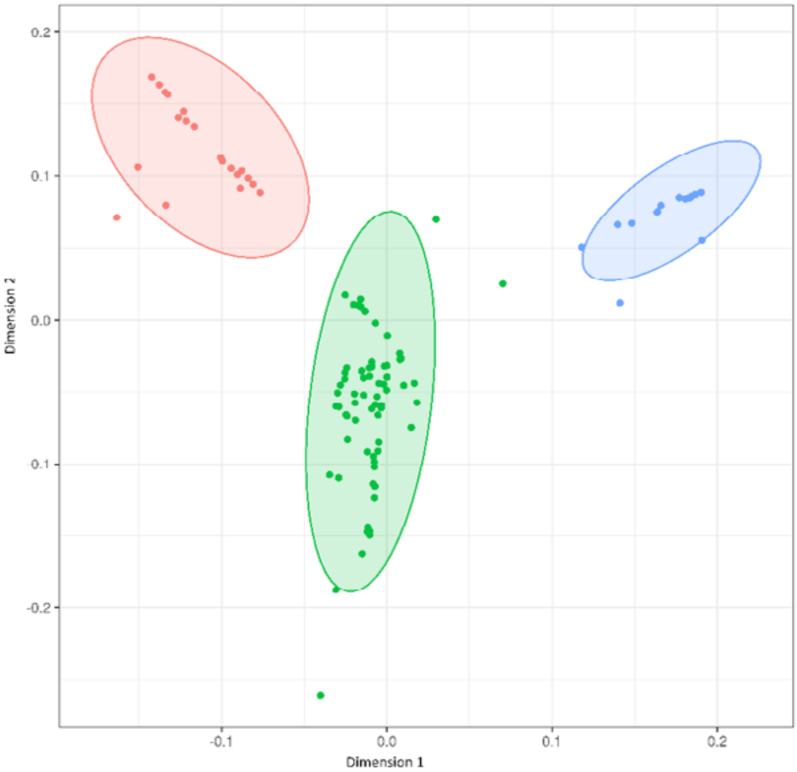


Figure 11. An example of a particular case in the dataset that may be produced by three users instead of two. A 2-dimensional MDS allows us to visualise three different regions in the domain space.

people to identify themselves when browsing, we will risk suffering a serious bias as mentioned in section II.

However, even if the algorithm performs worse than expected in real shared devices, it can still be improved to boost the accuracy. The following two techniques deserve further research:

- Instead of working at domain level in general, some non-discriminant domains could be split in subdomains. For instance, "facebook.com" would be a non-discriminant domain if both users visit Facebook. But considering "facebook.com/userA" and "facebook.com/userB" as different domains, a non-discriminant domain is transformed into two highly discriminant ones. It is the same strategy suggested to split hooks when two panellists share a domain (section IV, step 5). By applying similar pre-processing techniques to some popular domains, accuracy can be improved.

- Panellists could be asked to provide some additional information on how they browse the internet to improve the algorithm. Questions such as "Could you provide us some websites you usually visit that nobody else at your home does?" would improve accuracy and prevent a wrong identification of the hook. An alternative approach is to ask the user to identify herself in some initial browsing sessions. Although the user may alter the way he/she navigates in that session (e.g. avoiding sensible websites), the collected data could be enough to better train the algorithm. If people are required to identify their browsing sessions occasionally, for calibrating the algorithm, the passive nature of the data would not be compromised and churn rate would be limited.

Beyond its performance, the algorithm offers several advantages that facilitate its usage in production environments, compared to existing solutions.

- It is simple. It can be easily programmed in almost any programming language.

- It does not need to be executed in real time. In fact, it can be executed whenever it is needed, as it only requires (as input) the same data that is going to be processed. The only requirement is to have enough data stored to properly assign each domain to each user.

- The performance of the algorithm improves over time. The longer the period of time under analysis, the better the detection of similarities among domains and the better the accuracy. However, a limit should be set up to account for the possibility that users change their navigation habits. This topic requires further research.

Conclusions

We have shown that the way people use the internet is a personal trait that, in fact, should be considered as Personal Identifiable Information (PII). We can benefit from this fact to separate navigations of two or more users that are sharing the same metered browsing device.

To prove this concept, we have developed an algorithm that exploits the correlation among domains in browsing sessions. The algorithm was tested on a dataset created for this purpose, by mixing navigations of members of a Behavioural Panel (Netquest). We have reached 87.3% of accuracy in identifying website visits and 87.5% in identifying URLs visits.

This result opens a new line of research. There is plenty of room for improvement, as was presented in this paper.

Finally, we hope this work launches a debate about how we, as researchers, should approach new methodologies and data sources. The method we have explored to separate navigations is not perfect: some web visits, particularly those that are part of short browsing sessions, may be misclassified. This fact may cause discomfort in some researchers that prefer to work with the apparent certainty that self-reported data provides.

Despite the fact that self-reported data on online behaviours may be severely distorted, it is a safe place for researchers: if data makes no sense, just blame the online panel for not being able to recruit honest participants, without bearing in mind that is impossible to be honest when asked to report short and repetitive interactions that are repeated many times a day.

Behavioural data follows what may be called the “Heisenberg’s uncertainty principle on behavioural data”. Just like the original Heisenberg's principle stated that some pairs of physical properties cannot be known with unlimited precision, a perfect knowledge on online behaviours cannot be reached, while knowing all the surrounding circumstances without uncertainty. To know the latter, we need to ask. When asking, we alter behaviours and data is not passive anymore. New methods as the one presented in this paper offers a way to reduce the uncertainty around the behavioural data, without seeking to eliminate it completely.

Working with new types of data means taking risks; it means working with imperfect datasets, far away from the simplicity of survey data.

References

1. Baker, R., Brick, J.M., Bates, N.A., Battaglia. M., Couper, M.P., Dever, J.A., Gile, K.J., Tourangeau, R., (2013), “Summary report of the AAPOR Task Force on non-probability sampling”, Journal of Survey Statistics and Methodology, 2013;1:90–143.
2. Bishop, C. M (2006), “Pattern Recognition and Machine Learning”, ISBN-10: 0-387-31073-8.
3. Kuhn, M., Johnson K., (2010), “Applied Predictive Modeling”, ISBN 978-1-4614-6848-6.
4. Lozar-Manfreda, K. & Vehovar, V. (2008), “Internet surveys”, in E.D. de Leeuw, J.J. Hox & D.A. Dillman (Red.), International handbook of survey methodology. New York: Erlbaum.
5. ESOMAR (2016), “Global Market Research 2016”, ISBN: 92-831-0282-7.
6. Revilla, M., Ochoa, C., Loewe G., Voorend, R. (2015), “When should we ask, when should we measure? Comparing information from passive and active data collection”, ESOMAR CONGRESS 2015, ISBN: 92-831-0283-5.
7. Young, Forrest W. (1987). Multidimensional scaling: History, theory, and applications. Lawrence Erlbaum Associates. ISBN 978-0898596632.

Authors

Carlos Ochoa has an Engineering Degree in Telecommunications (UPC). He is experienced in consultancy, sales and product management. Formerly Operations Director at Netquest, he is currently in charge of defining and implementing the marketing strategy of the company as the Chief Client Officer, as well as fostering innovation projects in the quality data collection area. He has authored many papers and presentations at several market research events in his role of Innovation Leader at Netquest and as an active collaborator of the joint research program that Netquest has with the RECSM (Research and Expertise Centre for Survey Methodology, UPF, Barcelona). He is the main author of the recently published "Behavioural Data 101".

cochoa@netquest.com

Carlos Bort holds a MSc in Statistics and Operations Research, MSc in Actuarial Science and a BA in Economics. In his professional career he worked with data science through consultancy, start-ups and market research. He is an active member of the data-community, as co-organisator of Machine Learning Barcelona and BCN-R users group meetups. He has won several datathon competitions. In 2015, he joined Netquest to establish a data science department. The department grew to six data scientists where they continue to optimise the panel performance using cutting edge machine learning techniques. At the moment of the writing of this article, Carlos was the Lead Data Scientist at OKULANT and a frequent Netquest collaborator. Currently he is the Co-Founder & Data Scientist at xplore.ai.

carlosebort@gmail.com

Josep Miquel Porcar is part of the Data Science and Innovation Team at Netquest. He has a degree in Mathematics and holds a Master's Degree in Statistics and Operations Research with mention in Bioinformatics by the Universitat Politècnica de Catalunya. He developed his thesis on "NMF for deciphering latent signals of mutational processes". His main field of expertise is on machine learning related to consultancy and bioinformatics. Also, Josep Miquel is keen on music. He plays the saxophone and plays in the "Banda de la Societat Musical del País Valencià a Barcelona " orchestra, as well as participating in volunteering social activities.

mporcar@netquest.com

Trade Promotion Optimisation

TRANSFORMING PROMOTIONAL SPENDING FROM A COST OF DOING BUSINESS INTO A DRIVER OF GROWTH

Donald E. Schmidt, Ph.D.
Independent Business Analytics Consultant

Classifications, Key Words:

- Trade Promotions
- Analytics

Abstract

This paper describes an analytical methodology called Trade Promotion Optimisation (TPO). TPO is an advanced analytic that focuses on improving the ROI and sales productivity of trade funds spent against promotions. It focuses on three decision points: how to allocate spending among products, which vehicles to use, and how to optimise price discounting. I discuss the incremental sales and ROI benefits of promotion optimisation as it relates to improvements in the effectiveness of current spending. However, TPO is also a flexible methodology to determine whether current promotion funds can be reduced or, on the flipside, the implications of increasing promotion support. I provide examples of TPO outputs and discuss how this information can improve the productivity of promotional trade spending.

Why are Trade Promotion Tactics so Controversial

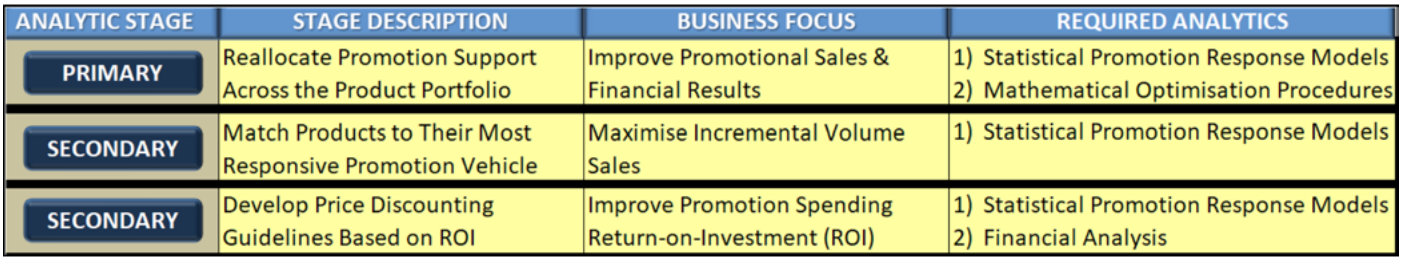
Trade promotions have become a common business practice for consumer-packaged goods (CPG) companies. They are a prevalent sales tactic because of their popularity among retailers and consumers alike. However, an analysis using data from a Catalina Marketing study (Rapperport, 2015) concluded that the top 100 CPG brands in the US were losing market share, attributable largely to the inefficiency of the trade promotions that they were running. Most promotions produce incremental volume sales. Still, poorly designed promotions or those run on unresponsive products can erode retail dollar sales and share. Hence, many managers are sceptical of the value of trade promotions for facilitating business growth. Surveys have found that most CPG executives question the practice. Many see it merely as a cost of doing business.

A detailed analysis by Andersen Consulting (Orler and Sotzing, 1997) concluded that trade promotions not only have limited effectiveness for business growth but introduced several inefficiencies into core business processes such as supply chain logistics and operations planning. Yet, they are mandated by retailer pressure and the belief that they are a “necessary evil” in a hypercompetitive marketplace.

Since companies frequently treat trade promotions as a cost of doing business, managers often neglect the analytics that could make them more effective. Trade Promotion Optimisation (TPO) was born out of the need to move trade promotions from a largely non-productive cost to a driver of business growth. Let's review these analytics and the value that they provide.

Components of trade promotion optimisation

There are three stages in the TPO process, coinciding with different promotion decision points. We complete these analyses for retail accounts, the geographic level of promotion planning and execution. I have described these three stages in **Table 1**.



ANALYTIC STAGE	STAGE DESCRIPTION	BUSINESS FOCUS	REQUIRED ANALYTICS
PRIMARY	Reallocate Promotion Support Across the Product Portfolio	Improve Promotional Sales & Financial Results	1) Statistical Promotion Response Models 2) Mathematical Optimisation Procedures
SECONDARY	Match Products to Their Most Responsive Promotion Vehicle	Maximise Incremental Volume Sales	1) Statistical Promotion Response Models
SECONDARY	Develop Price Discounting Guidelines Based on ROI	Improve Promotion Spending Return-on-Investment (ROI)	1) Statistical Promotion Response Models 2) Financial Analysis

Table 1. Three Analytic Stages of Trade Promotion Optimisation.

The Reallocation of Promotion Support

The first and most important step in TPO is the reallocation of promotion support, product-by-product across the portfolio. This reallocation is relative to current promotion execution levels reported on a syndicated database. Syndicated databases report promotion execution, such as average weeks of promotion, or percent all-commodity volume (%ACV) promoted that have occurred within a specified timeframe. TPO takes this beginning point over the most recent year of data and recommends the "ideal."

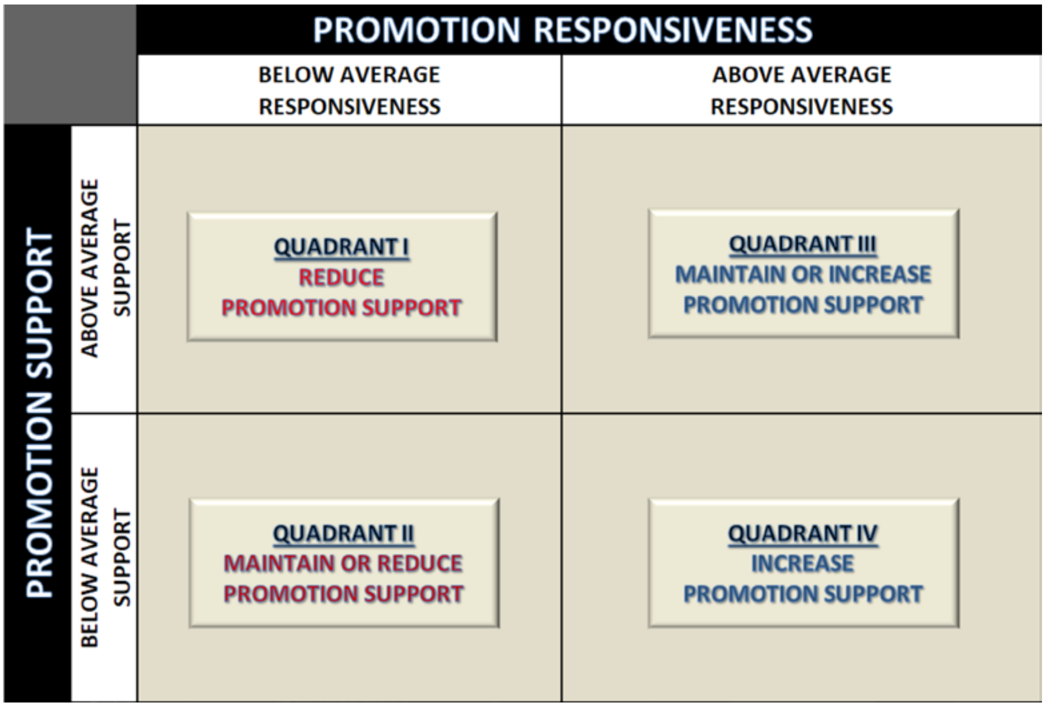
This "ideal" promotion allocation scheme is determined in respect to the specific business goals underlying promotional spending. Product-by-product reallocation recommendations will vary according to the specific sales or financial objectives. Examples of business goals related to promotional support are:

- Maximise incremental volume sales, aggregated across the product portfolio.
- Maximise incremental retail dollar sales, aggregated across the portfolio.
- Maximise the Return-on-Investment (ROI) of the total promotional spend.

Figure 1 illustrates an intuitive logic for promotion support reallocations. Start by classifying each product in the portfolio into one of four quadrants based on two dimensions. The first dimension is the relative level of promotion support currently (high vs. low) determined from syndicated data.

The second dimension is the relative incremental responsiveness of each product-item to promotional support. The bifurcation of products based on the relative responsiveness to promotional support is based on a custom calculation. I use the promotion coefficients and discounting elasticities from a statistical response model to derive a single response variable. The question addressed by this custom calculation is "if each product ran the same mix of promotion vehicles and the same discounts, what is the relative responsiveness among the products in the portfolio?" The promotion, mix and discounts used are based on current executed averages from syndicated data. This provides an "apples-to-apples" comparison of product responsiveness. Split products into high vs. low promotion response groups using this response measure.

The logic behind this analysis is straightforward. Products that have high levels of promotion support but are unresponsive to this support (Quadrant I) are candidates for reduced support. On the other hand, products that have low levels of promotion support, but are highly responsive to this support (Quadrant IV) are candidates for increased support. So, shift promotion support from Quadrant I to Quadrant IV.



		PROMOTION RESPONSIVENESS	
		BELOW AVERAGE RESPONSIVENESS	ABOVE AVERAGE RESPONSIVENESS
PROMOTION SUPPORT	ABOVE AVERAGE SUPPORT	QUADRANT I REDUCE PROMOTION SUPPORT	QUADRANT III MAINTAIN OR INCREASE PROMOTION SUPPORT
	BELOW AVERAGE SUPPORT	QUADRANT II MAINTAIN OR REDUCE PROMOTION SUPPORT	QUADRANT IV INCREASE PROMOTION SUPPORT

Figure 1. Analytical Logic for Reallocating Promotion Support

Figure 1 offers a good conceptual picture of what we are trying to accomplish with promotion reallocation. However, from an analytical viewpoint, the procedures involved are much more complex. When you consider: multiple goals (sales and financial), numerous products, the need to provide concrete recommendations for each, and constraints; the computational load becomes untenable and requires automation. Additionally, the reallocation process involves a complex array of data linkages and technical procedures.

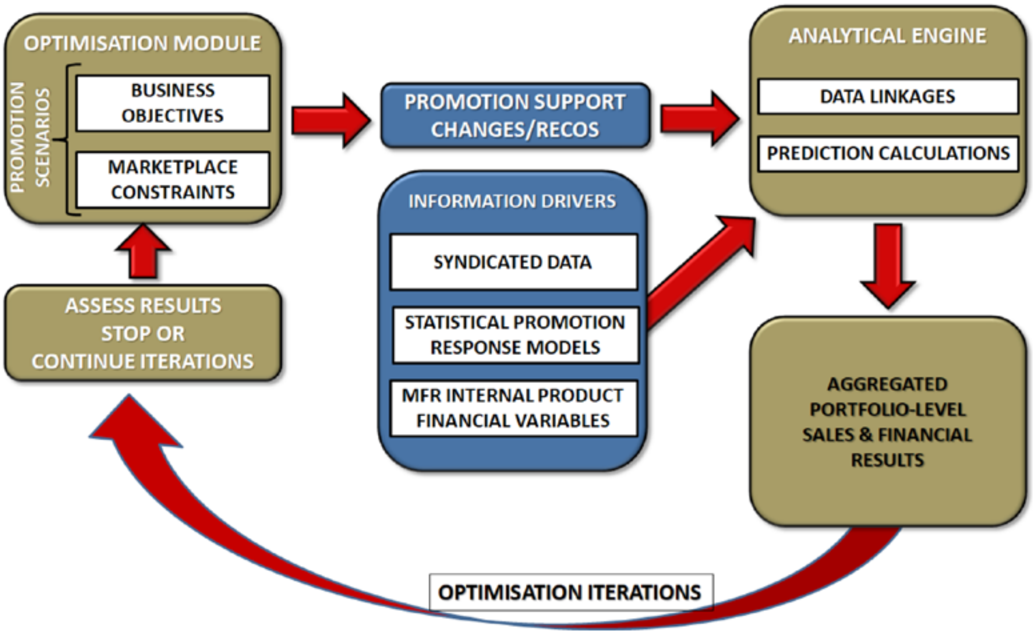

OPTIMISATION MODULE
PROMOTION SCENARIOS
BUSINESS OBJECTIVES
MARKETPLACE CONSTRAINTS
PROMOTION SUPPORT CHANGES/RECOS
ANALYTICAL ENGINE
DATA LINKAGES
PREDICTION CALCULATIONS
INFORMATION DRIVERS
SYNDICATED DATA
STATISTICAL PROMOTION RESPONSE MODELS
MFR INTERNAL PRODUCT FINANCIAL VARIABLES
ASSESS RESULTS STOP OR CONTINUE ITERATIONS
AGGREGATED PORTFOLIO-LEVEL SALES & FINANCIAL RESULTS
OPTIMISATION ITERATIONS

Figure 2. Promotion Optimisation Process

Figure 2 is a process map for promotion optimisation that illustrates these complexities.

Here are the major steps in the promotion reallocation optimisation process:

- **Optimisation Setup:** Linear programming (LP) is the preferred optimisation procedure because of its widespread availability and applicability for the task. The first step is to define three elements of a promotion scenario: 1) The primary goal of the optimisation, 2) Means to the goal, and 3) Constraints. The primary goal is the most important outcome you are trying to achieve (e.g. maximise incremental promoted sales). The means involve the variable that you will change to achieve the goal. In the case of promotion reallocation, this would involve product-item level changes in promotion support. Finally, constraints can involve two elements: 1) limits on the means – for example, how much of a promotion change you will allow for each product relative to current support, or 2) secondary goals – for example, spending ROI targets (i.e. ROI improvements could be a secondary goal specified as a constraint). Once we launch the optimiser with the defined scenario, an iterative process begins, changing the “means” variable to improve the primary goal outcomes within the constraints.

- **Evaluation of Recommendations Using Statistical Models:** I use statistical response models to evaluate recommendations associated with each optimisation iteration. The models I have used successfully are proprietary to Nielsen or IRI, the two major CPG data and analytics vendors. These are multiplicative time series models based on weekly store-level data. The modelled dependent variable is changes in retail volume sales. These are integrated models that include both regular pricing and promotional variables. For TPO, I use only the promotion results, which are regression weights for vehicle flags and promoted price elasticities modelled as a log-linear variable. Vehicle flags are dummy variables, coded 1 or 0 for the presence or absence of a feature ad or display respectively during a given data week. They allow modelers to calculate the average sales multiplier associated with a feature ad, display, or combination. The functional model form is illustrated in **Figure 3**. The regular pricing component includes log-linear elasticity and cross-elasticity coefficients quantifying the impact of the retail price point and relevant competitive item price gaps. The control variables provide covariate adjustments for what are ostensibly nuisance variables that can affect the model estimates and statistical errors. Control variables are used to adjust the results, but not for any predictive purposes. The regular price coefficients can be used to optimise regular pricing, a sibling analysis to TPO (see Schmidt, 2017).

- **Analytical Engine:** The models take the optimiser promotion support recommendations at each iteration and predict resulting changes in product-item retail volume sales. We combine these predictions with syndicated price, promotion, and sales data, along with financial measures such as product-item unit margins and list prices to calculate additional sales and financial measures. Examples of these added measures are changes in incremental retail dollar sales, manufacturer revenues, and ROI.

- **Portfolio Assessment:** All of the calculations performed within the analytical engine are at the product-item level of detail. The next step is to aggregate these results to the portfolio level, because the primary goal specified in

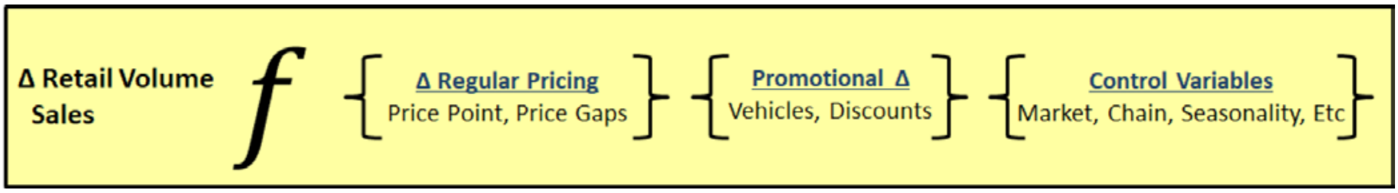


Figure 3. Promotion Optimisation Process

the optimisation scenario is the total across all products.

- **Linear Programming Iterations:** LP involves an iterative process, with each successive iteration producing a better solution relative to the primary goal within the defined constraints. These iterations continue until no further improvements, within a specified precision limit, can be gained.

These optimisation procedures are flexible and allow you to evaluate several promotion scenario variations. Certainly, you can change goals and constraints. However, you also can go beyond these simple scenario variants and change the underlying business questions behind an optimisation. For example:

- How should you reallocate current promotion support to improve sales productivity and ROI?

- Can you reduce current spending and still maintain current incremental sales?

- What promotion spending is required to attain some predetermined sales or financial targets?

These are common sales management questions and are within the scope of this optimisation technology.

The reallocation optimisation process produces two classes of outputs:

- **Reallocation Recommendations:** We can compare the optimisation recommendations to current product promotion allocations at the product-item level of detail.

- **Aggregated Portfolio Results:** The results of the optimisation aggregated across the product portfolio. These are results based on key performance measures and ROI calculations.

Table 2 is a simplified example of the outputs derived from the reallocation optimisation process.

Optimising Promotion Vehicle Selection

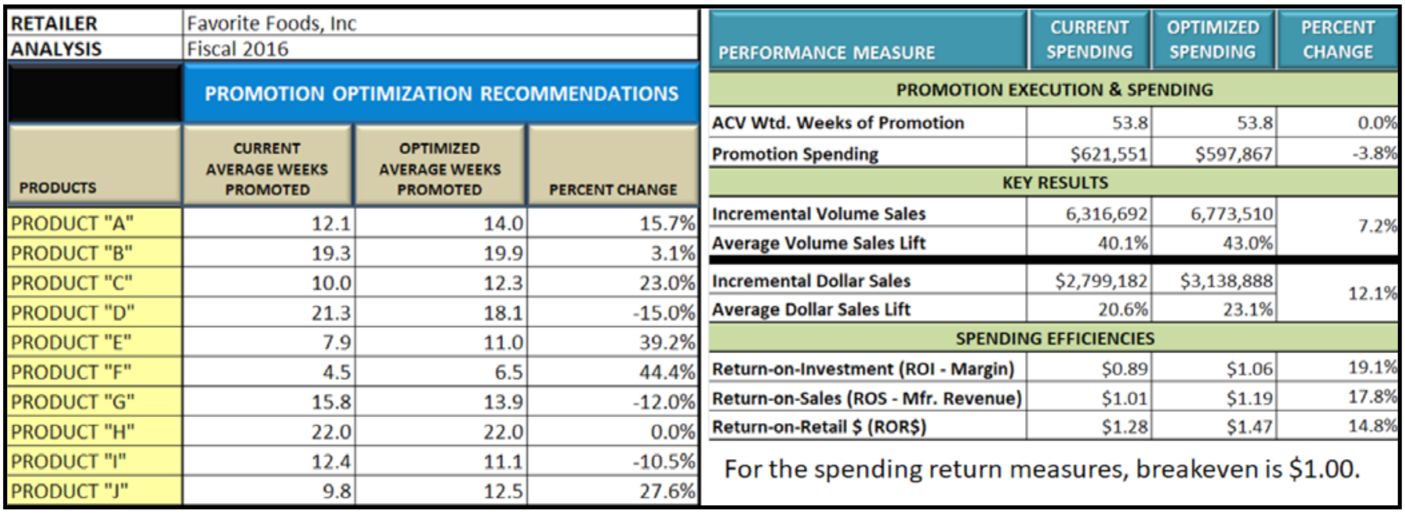
The next step in the process is to match products to the most responsive promotion vehicle. Trade promotions are retailer initiated, typically paid for by manufacturers. The primary promotion vehicles are:

- Retailer-produced and placed feature ads.

- In-store product displays.

- A combination of feature ads and displays.

- Temporary price reductions (TPRs), executed either alone or in conjunction with the other vehicles.



RETAILER	Favorite Foods, Inc		
ANALYSIS	Fiscal 2016		
	PROMOTION OPTIMIZATION RECOMMENDATIONS		
PRODUCTS	CURRENT AVERAGE WEEKS PROMOTED	OPTIMIZED AVERAGE WEEKS PROMOTED	PERCENT CHANGE
PRODUCT "A"	12.1	14.0	15.7%
PRODUCT "B"	19.3	19.9	3.1%
PRODUCT "C"	10.0	12.3	23.0%
PRODUCT "D"	21.3	18.1	-15.0%
PRODUCT "E"	7.9	11.0	39.2%
PRODUCT "F"	4.5	6.5	44.4%
PRODUCT "G"	15.8	13.9	-12.0%
PRODUCT "H"	22.0	22.0	0.0%
PRODUCT "I"	12.4	11.1	-10.5%
PRODUCT "J"	9.8	12.5	27.6%

PERFORMANCE MEASURE	CURRENT SPENDING	OPTIMIZED SPENDING	PERCENT CHANGE
PROMOTION EXECUTION & SPENDING			
ACV Wtd. Weeks of Promotion	53.8	53.8	0.0%
Promotion Spending	$621,551	$597,867	-3.8%
KEY RESULTS			
Incremental Volume Sales	6,316,692	6,773,510	7.2%
Average Volume Sales Lift	40.1%	43.0%	
Incremental Dollar Sales	$2,799,182	$3,138,888	12.1%
Average Dollar Sales Lift	20.6%	23.1%	
SPENDING EFFICIENCIES			
Return-on-Investment (ROI - Margin)	$0.89	$1.06	19.1%
Return-on-Sales (ROS - Mfr. Revenue)	$1.01	$1.19	17.8%
Return-on-Retail $ (ROR$)	$1.28	$1.47	14.8%

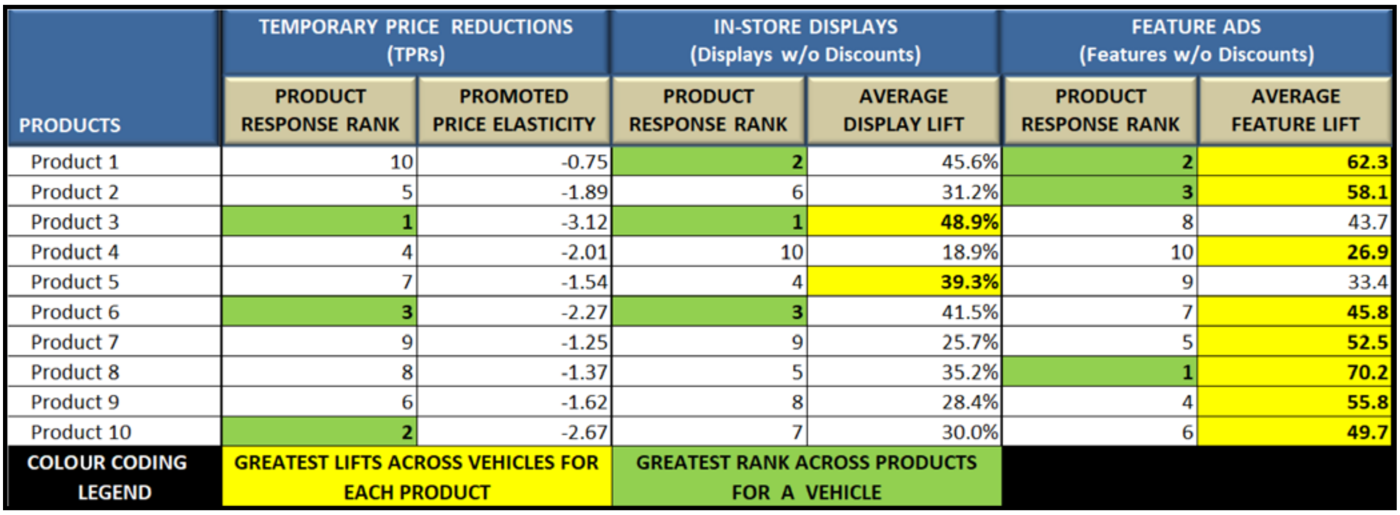
For the spending return measures, breakeven is $1.00.

Table 2. Promotion Optimisation – Support Reallocation Outputs

Manufacturers can participate in feature ads and display promotions without a discount, but this is relatively rare because retailers balk at non-discounted promotions. Promotional “space” is limited and retailers want to ensure the promotions that they run are attractive to consumers. This usually requires a discount.

The primary sources of information for promotion vehicle decisions involves lift coefficients and elasticities from statistical response models. We can associate promotion vehicles to products in two ways:

- **Within Product Analysis:** For any given product, which promotion vehicle provides the greatest lift? If you want to promote a specific product-item, which promotion vehicle is the best?
- **Across Product Analysis:** Across all products, which are the most responsive to feature ads, in-store displays, or TPRs? If you can participate in a feature or display promotion, which products are best?



PRODUCTS	TEMPORARY PRICE REDUCTIONS (TPRs)		IN-STORE DISPLAYS (Displays w/o Discounts)		FEATURE ADS (Features w/o Discounts)	
	PRODUCT RESPONSE RANK	PROMOTED PRICE ELASTICITY	PRODUCT RESPONSE RANK	AVERAGE DISPLAY LIFT	PRODUCT RESPONSE RANK	AVERAGE FEATURE LIFT
Product 1	10	-0.75	**2**	45.6%	**2**	**62.3**
Product 2	5	-1.89	6	31.2%	**3**	**58.1**
Product 3	**1**	-3.12	**1**	**48.9%**	8	43.7
Product 4	4	-2.01	10	18.9%	10	**26.9**
Product 5	7	-1.54	4	**39.3%**	9	33.4
Product 6	**3**	-2.27	**3**	41.5%	7	**45.8**
Product 7	9	-1.25	9	25.7%	5	**52.5**
Product 8	8	-1.37	5	35.2%	**1**	**70.2**
Product 9	6	-1.62	8	28.4%	4	**55.8**
Product 10	**2**	-2.67	7	30.0%	6	**49.7**
COLOUR CODING LEGEND	GREATEST LIFTS ACROSS VEHICLES FOR EACH PRODUCT		GREATEST RANK ACROSS PRODUCTS FOR A VEHICLE			

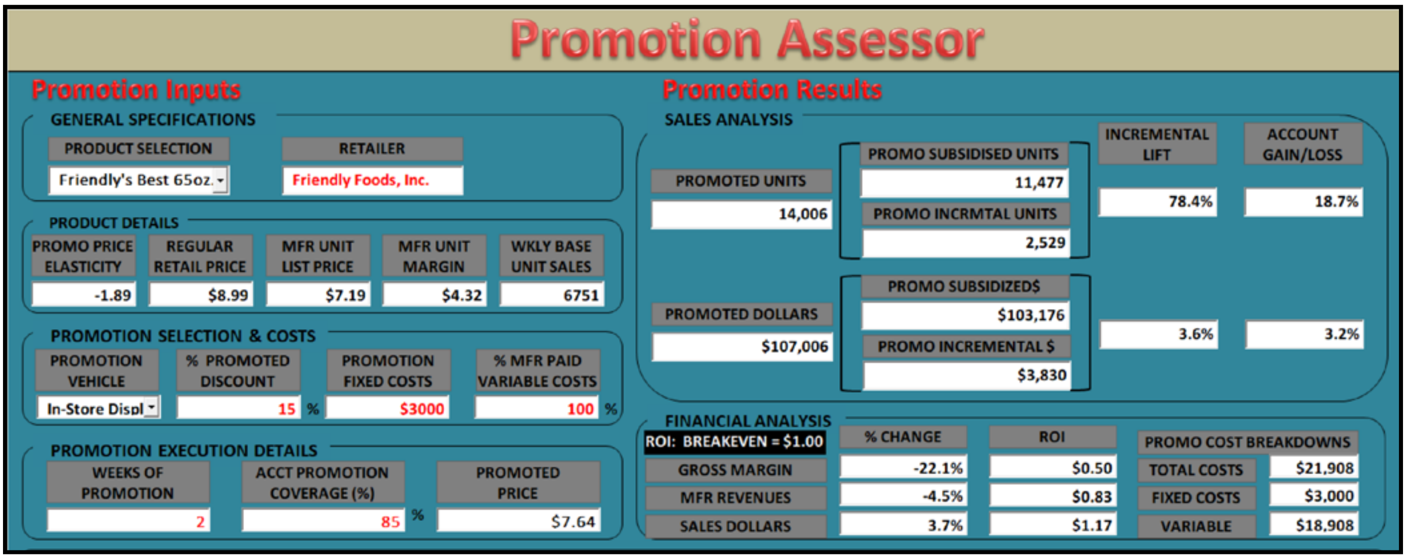
Table 3. Analysis & Outputs for Vehicle Selection Optimisation

Table 3 is an example of the outputs for the vehicle selection analysis. Yellow shading represents the within product comparison. Green shading represents the results of the cross-product analysis.

However, it may not be as simple as comparing lift coefficients. The analysis in Table 3 looks at each of the promotion options in isolation. However, a typical trade promotion is a combination of vehicles and discounts. We need to assess the combination beyond the individual elements. We can take a specific promotion combination and estimate expected incrementality and financial results. This analysis utilises the results from promotion response models, product financial information, and syndicated sales data. **Figure 4** illustrates what an analysis integrating vehicles and discounting would look like. The example is a two-week in-store product display with a 15% discount run in 85% of the retailer’s stores. This is a typical execution.

Discounting Guidelines

The final component of TPO is the development of discounting guidelines based on ROI. The two variables that are strong drivers of promotional ROI are costs and responsiveness. Costs are determined by the depth of the price discount against the non-promoted price. We quantify responsiveness using promoted price elasticities from the statistical models. To illustrate these relationships, I ran a series of simulations. I evaluated a range of elasticities holding regular price and base margin constant. **Figure 5** displays the results. ROI declines as the discount increases when elasticity is high. With high elasticities, ROI remains positive across a wide range of discounts. For low elasticities, ROI’s are significantly below breakeven at all discounts.



Promotion Assessor

Promotion Inputs

GENERAL SPECIFICATIONS

PRODUCT SELECTION	RETAILER
Friendly's Best 65oz.	Friendly Foods, Inc.

PRODUCT DETAILS

PROMO PRICE ELASTICITY	REGULAR RETAIL PRICE	MFR UNIT LIST PRICE	MFR UNIT MARGIN	WKLY BASE UNIT SALES
-1.89	$8.99	$7.19	$4.32	6751

PROMOTION SELECTION & COSTS

PROMOTION VEHICLE	% PROMOTED DISCOUNT	PROMOTION FIXED COSTS	% MFR PAID VARIABLE COSTS
In-Store Displ	15 %	$3000	100 %

PROMOTION EXECUTION DETAILS

WEEKS OF PROMOTION	ACCT PROMOTION COVERAGE (%)	PROMOTED PRICE
2	85 %	$7.64

Promotion Results

SALES ANALYSIS

PROMOTED UNITS	PROMO SUBSIDISED UNITS	PROMO INCRMTAL UNITS	INCREMENTAL LIFT	ACCOUNT GAIN/LOSS
14,006	11,477	2,529	78.4%	18.7%
PROMOTED DOLLARS	PROMO SUBSIDIZED$	PROMO INCREMENTAL $		
$107,006	$103,176	$3,830	3.6%	3.2%

FINANCIAL ANALYSIS

ROI: BREAKEVEN = $1.00	% CHANGE	ROI	PROMO COST BREAKDOWNS	
GROSS MARGIN	-22.1%	$0.50	TOTAL COSTS	$21,908
MFR REVENUES	-4.5%	$0.83	FIXED COSTS	$3,000
SALES DOLLARS	3.7%	$1.17	VARIABLE	$18,908

Figure 4. Promotion Simulator

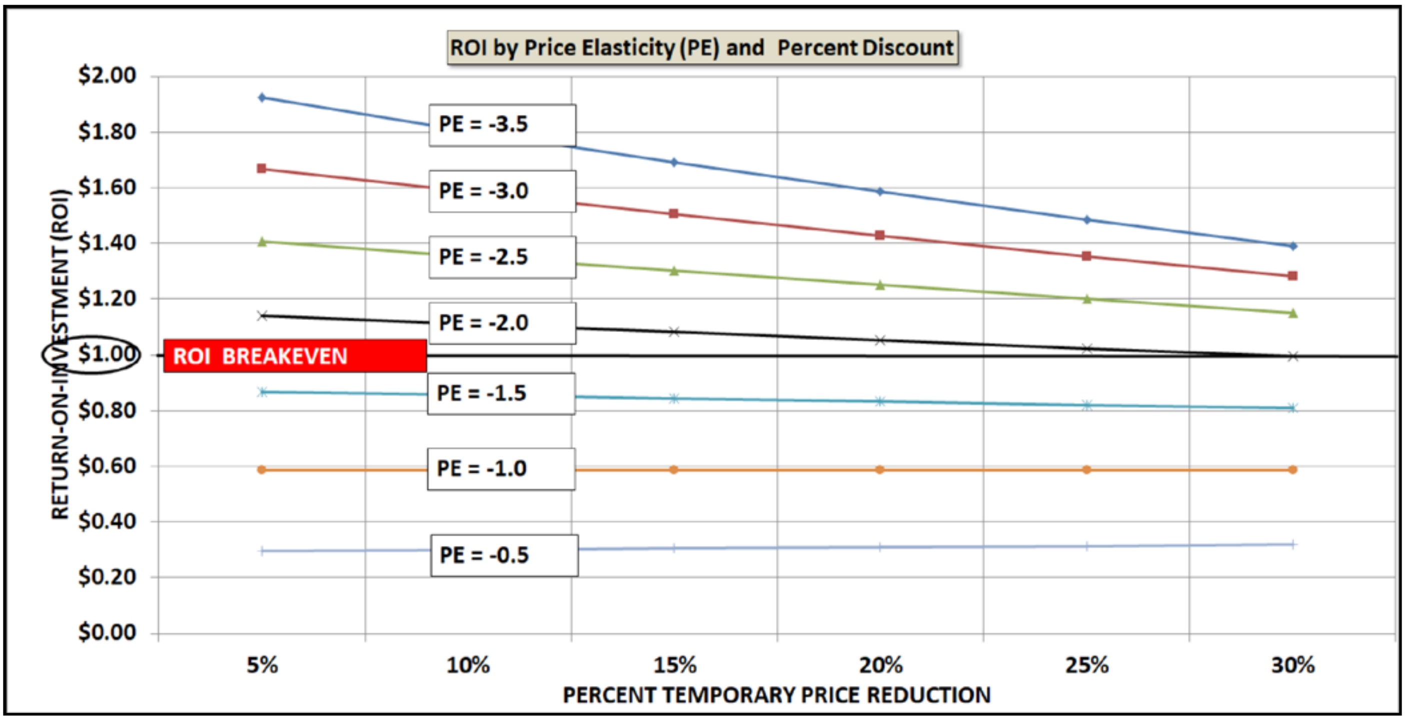
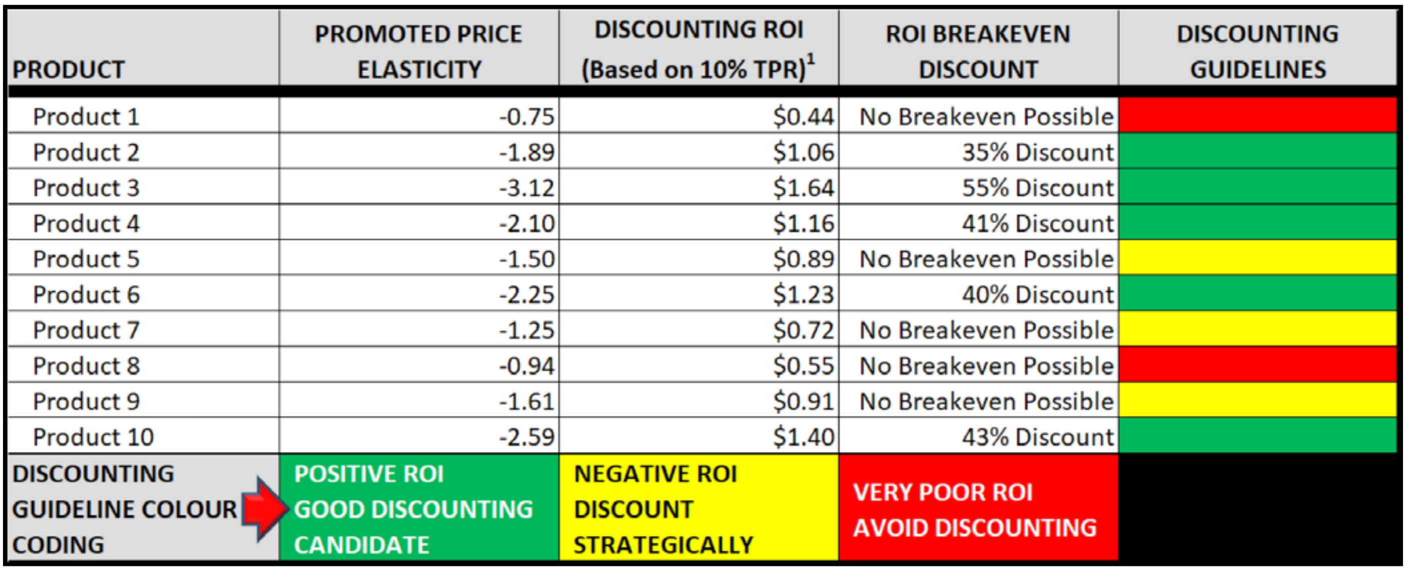


Figure 5. Promoted Price Elasticity Determines Discounting ROI and the Breakeven Point

These results allow us to formulate discounting guidelines focused on ROI. Using promoted price elasticities, we can calculate the breakeven discount. Of course, for some products with relatively low-price elasticities, breakeven ROI may not be possible at any level. However, we can readily determine this.

We can classify products according to how good or bad the ROI result will be at some average discount value. Table 4 shows the discounting ROI for a 10% TPR across the ten products in this example. We can determine for each product-item in the portfolio, whether price reductions can be aggressive, whether we should limit discounts, or whether we should avoid discounting altogether, depending on the breakeven point. Various strategic or tactical considerations also may come into play when determining discounting guidelines for some products, but the analytics are a good beginning reference point.



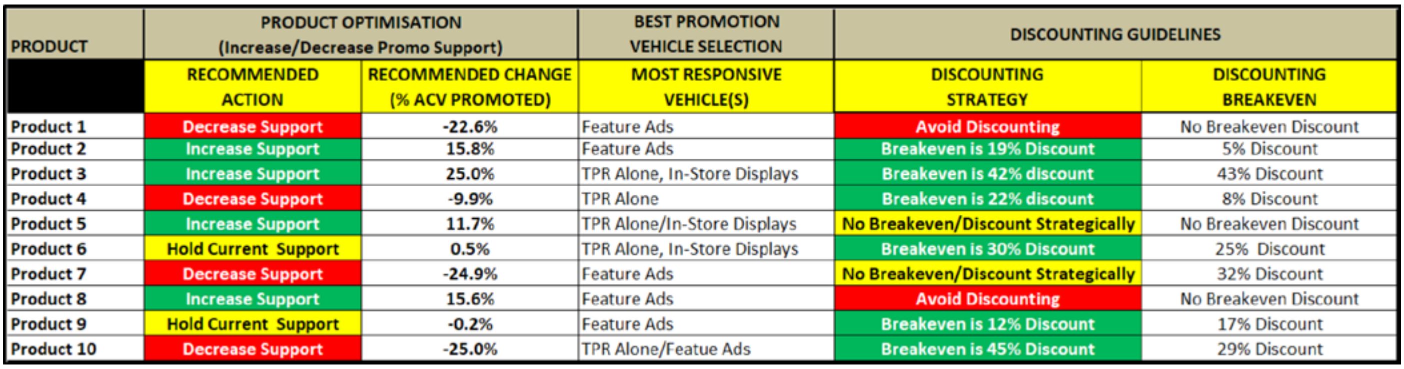
PRODUCT	PROMOTED PRICE ELASTICITY	DISCOUNTING ROI (Based on 10% TPR)[1]	ROI BREAKEVEN DISCOUNT	DISCOUNTING GUIDELINES
Product 1	-0.75	$0.44	No Breakeven Possible	🟥
Product 2	-1.89	$1.06	35% Discount	🟩
Product 3	-3.12	$1.64	55% Discount	🟩
Product 4	-2.10	$1.16	41% Discount	🟩
Product 5	-1.50	$0.89	No Breakeven Possible	🟨
Product 6	-2.25	$1.23	40% Discount	🟩
Product 7	-1.25	$0.72	No Breakeven Possible	🟨
Product 8	-0.94	$0.55	No Breakeven Possible	🟥
Product 9	-1.61	$0.91	No Breakeven Possible	🟨
Product 10	-2.59	$1.40	43% Discount	🟩
DISCOUNTING GUIDELINE COLOUR CODING ➡	POSITIVE ROI GOOD DISCOUNTING CANDIDATE	NEGATIVE ROI DISCOUNT STRATEGICALLY	VERY POOR ROI AVOID DISCOUNTING	

Table 4. Discounting ROI and Discounting Guidelines

Integrating the Separate TPO Analytics

In this final step, we integrate the results from the three TPO components: promotion support reallocation, vehicle selection, and discounting guidelines. We want to form a comprehensive picture of how to maximise return on promotional spending. **Table 5** provides an example of this integration. If you are a company that competes in several categories and has a large number of product offerings, this list may be expansive. It is important to colour code the table to provide a quick visual overview of the results.

The most impactful analysis for TPO is the product support reallocations. This is the foundation for improving promotional spending ROI. Vehicle selection and discounting guidelines are subsidiary analyses that can help you tweak the plan and gain additional efficiencies.



PRODUCT	PRODUCT OPTIMISATION (Increase/Decrease Promo Support)		BEST PROMOTION VEHICLE SELECTION	DISCOUNTING GUIDELINES	
	RECOMMENDED ACTION	RECOMMENDED CHANGE (% ACV PROMOTED)	MOST RESPONSIVE VEHICLE(S)	DISCOUNTING STRATEGY	DISCOUNTING BREAKEVEN
Product 1	Decrease Support	-22.6%	Feature Ads	Avoid Discounting	No Breakeven Discount
Product 2	Increase Support	15.8%	Feature Ads	Breakeven is 19% Discount	5% Discount
Product 3	Increase Support	25.0%	TPR Alone, In-Store Displays	Breakeven is 42% discount	43% Discount
Product 4	Decrease Support	-9.9%	TPR Alone	Breakeven is 22% discount	8% Discount
Product 5	Increase Support	11.7%	TPR Alone/In-Store Displays	No Breakeven/Discount Strategically	No Breakeven Discount
Product 6	Hold Current Support	0.5%	TPR Alone, In-Store Displays	Breakeven is 30% Discount	25% Discount
Product 7	Decrease Support	-24.9%	Feature Ads	No Breakeven/Discount Strategically	32% Discount
Product 8	Increase Support	15.6%	Feature Ads	Avoid Discounting	No Breakeven Discount
Product 9	Hold Current Support	-0.2%	Feature Ads	Breakeven is 12% Discount	17% Discount
Product 10	Decrease Support	-25.0%	TPR Alone/Featue Ads	Breakeven is 45% Discount	29% Discount

Table 5. Integrated TPO Analytics

Summary and Conclusions

Trade promotions are a common retail tactic for CPG companies. However, many managers feel that this spending is cost-ineffective and produces a poor return. It is viewed largely as a non-productive cost, rather than as a sound business practice that can be used to drive sales and financial growth. In this article, I have introduced Trade Promotion Optimisation (TPO) as an analytical solution to address these shortcomings.

The primary focus of TPO is to improve the sales productivity and ROI of promotional spending. We do this by improving three key promotion decisions: the allocation of spending among products, selection of responsive promotion vehicles, and effective temporary price discounting. These are separate decisions and each requires a different analytical approach. However, by integrating the results from the three, we can renovate what is perceived to be inefficient spending to an effective component of business growth.

TPO uses three analytical components: statistical response modelling, mathematical optimisation procedures, and financial analyses linked to the modelling and optimisation work. The underlying theme for these analyses is the leveraging of promotion response information as the backbone of an effective trade promotion plan.

In a hypercompetitive marketplace, the advantage goes to those who can effectively leverage the resources available to them. Use Trade Promotion Optimisation analytics to facilitate better spending decisions and to forge a pathway to improved business growth.

References

1. Orler, Victor & Sotzing, Steve, The Daunting Dilemma of Trade Promotion. Anderson Consulting Publication, 1997.

2. Rapperport, Jamie, How Ineffective Promotions are Dragging Down CPG Brands. Progressive Grocer, October 15, 2015.

3. Schmidt, Donald, How to Achieve Effective Pricing: Leveraging the Power of Price Optimisation. DMA Analytics Journal, 2017.

Author

Donald E. Schmidt, Ph.D. is a 35-year marketing research veteran specializing in modeling and advanced analytics. He has worked for Quaker Oats, A. C. Nielsen, R. J. Reynolds, and Nestle Purina PetCare. Don's expertise includes pricing, promotion, and trade spending analytics and optimisation, marketing plan effectiveness, and special analytics focused against marketing, sales, and marketplace issues. While in the Nestle organisation, he was instrumental in developing and launching a Global Price Optimisation capability. Don is the author of the book, "The Pricing Doctor Is In: A Veteran Pricing Analyst Reveals an Innovative Approach to Effective Retail Pricing".

datadoc2@live.com

Consumer Journey: A New Perspective to the Attribution Problem

Shawn Song, Ph.D.
PHD Media

Charlotte Ma, MS
SapientRazorfish

Pranav Patil, MBA
Annalect

Classifications, Key Words:

- Attribution Modelling
- Consumer Journey
- Effectiveness Measurement
- KPI
- Optimisation

Abstract

Using an innovative approach to the attribution problem, this study showed insights regarding advertising impacts on consumer journey, illustrated in a case of a furniture retailer. Guided by behaviour science, this study shifted the focus of the attribution problem from a single action (e.g., purchase) to an experience (e.g., consumer journey). The innovative approach showed advantages over conventional methods. Results of the study challenged the validity of commonly adopted media measurement and optimisation practice.

Introduction

Effectiveness measurement is a central problem for important decisions in advertising management. This problem can be complex because advertising campaigns often include multiple media channels and tactics designed with different objectives. Commonly used methods include attribution modelling, which uses machine learning or statistical methods to attribute credit to individual advertising exposures based on individual (cookie) level data of advertising exposures and conversion events (e.g., purchase).

The current study approached the attribution problem with a different perspective. As opposed to focusing on a single event of conversion, we regarded consumer purchase decision as a process of multiple stages. As consumers progress through the process or journey, their tasks change and their needs for information migrate accordingly. Advertising as a source of information may thus show varying impact at different stages of the journey.

Commonly used attribution approaches, which draw direct lines between advertising exposures and conversions, over-simplify the reality and can miss important insights. Our approach focuses on advertising impact on progression of consumer journey. By breaking down the attribution problem by stages of the consumer journey, the approach gives advertisers an ability to understand which media channels, at what point influenced the consumer.

We illustrated the approach using data provided by a major furniture retailer in the US. The data came from the retailer's

Data Management Platform, which tracks individual level information regarding advertising exposures, site activities, transactions, and consumer attributes (e.g., demographics and psychographics). The advertising exposures were recorded at quite granular level reflecting the digital tactics and channels utilised in their advertising campaigns.

The analysis has a few findings that may not be visible via traditional methods. It showed where a consumer in the journey is strongly influenced by her receptivity to advertising messages. Interestingly enough, we saw advertising made much greater impact earlier on, rather than in the later stages of the journey, which aligned with the shifted information needs of consumers. The results also showed problems of commonly used simplistic performance metrics such as conversion rate. If not used carefully, conversion rate can mislead an advertiser to non-optimal decisions in fund allocation.

Consumer Decision Process

Consumer decision processes have been extensively discussed in marketing literature although rarely seen in discussions of advertising media practice. With slight variation, theories generally describe the process as one that follows a few consecutive stages such as problem recognition, research, evaluation, purchase, loyalty etc (Court et al, 2009, Edelman and Singer, 2015, Kotler, 2012).

Furniture purchase can be a quite lengthy process given relatively high cost and social visibility of furniture products. It makes sense that consumers go through distinctive stages weighing different factors at different stages. The conceptual consumer journey framework in the study was developed by the retailer we worked with. This framework includes four conceptual stages including awareness of brand and its offerings, exploring ideas and product options, evaluating and comparing product items, and making a purchase. For simplicity, these are labelled as Awareness, Exploration, Evaluation, and Purchase in the rest of the paper.

The Data

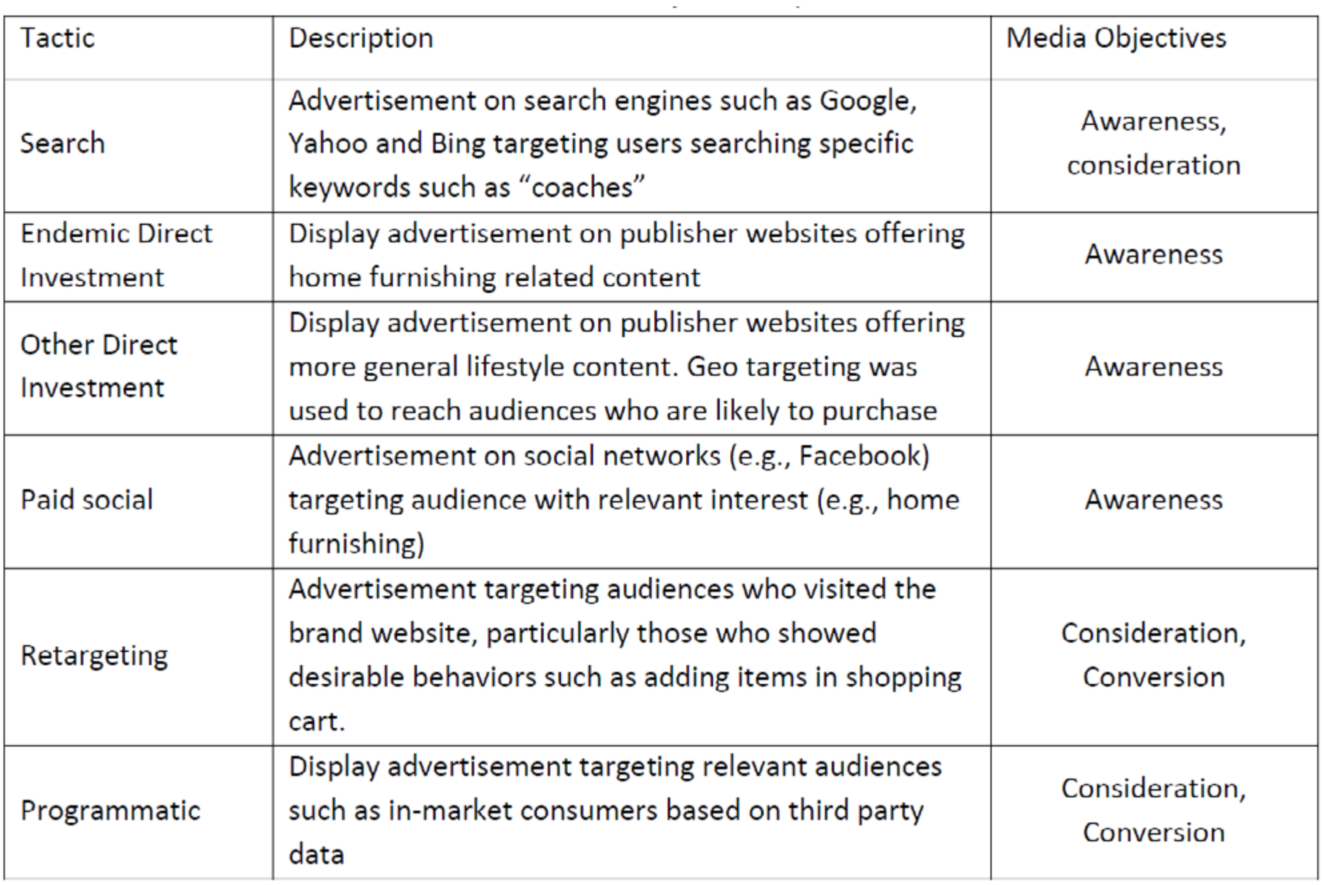
As mentioned earlier, this study utilised data from the retailer's Data Management Platform (DMP), which tracks advertising exposures, site behaviours, and purchases at individual (cookie) level. DMP is a recent advancement in advertising technology that gives advertisers visibility into the audiences their campaigns reach and an ability to track performance across digital channels. In this analysis, we selected a period of 90 days when the retailers ran a multi-channel advertising campaign. The digital media channels included paid search, social, endemic direct investment, other direct investment, programmatic, and site retargeting. Descriptions of these channels are shown in **Table 1**.

Milestones of the Consumer Journey

The conceptual consumer journey framework was operationalised into behavioural milestones using website visits recorded in the DMP. Essentially the visits were classified into the four stages of the journey based on the behavioural pattern within these visits. The retailer's website has been extensively used by their customers for research purpose; over 90% of their customers visited the website at some point prior to purchase. The behavioural pattern of these visits was very diverse in terms of number of pages viewed and types of pages viewed, signalling different purposes for these visits.

The classification was determined by judgement of three groups of experts who have expertise within the subject matter. Included were the marketing leaders of the retailer, their advertising agency strategy leads who managed the advertising campaign, and the researchers of the current study. Work sessions were conducted among these groups of experts to thoroughly discuss the meaning of these stages and gain alignment on specific site visits assigned to each stage.

The site visits were assigned to the four stages in the following manner. The Awareness stage



Tactic	Description	Media Objectives
Search	Advertisement on search engines such as Google, Yahoo and Bing targeting users searching specific keywords such as "coaches"	Awareness, consideration
Endemic Direct Investment	Display advertisement on publisher websites offering home furnishing related content	Awareness
Other Direct Investment	Display advertisement on publisher websites offering more general lifestyle content. Geo targeting was used to reach audiences who are likely to purchase	Awareness
Paid social	Advertisement on social networks (e.g., Facebook) targeting audience with relevant interest (e.g., home furnishing)	Awareness
Retargeting	Advertisement targeting audiences who visited the brand website, particularly those who showed desirable behaviors such as adding items in shopping cart.	Consideration, Conversion
Programmatic	Display advertisement targeting relevant audiences such as in-market consumers based on third party data	Consideration, Conversion

Table 1. Media objectives by tactics

included site visits that had three or fewer page views. The Exploration stage included visits that were either greater than three-page views or were repeated visitors. The Evaluation stage included visits that showed viewing of greater than three specific product items or specific e-commerce related activities such as shopping cart were viewed. And lastly, the Purchase stage was indicated by a transaction either online or offline.

Logistic Regression

We determined media impact on the consumer journey using a series of logistic regression models. A short description of logistic regression is included in the Appendix. In these models, the number of media exposures by media tactic were used as the predictive variables and progression of the journey (e.g., individuals who proceeded to subsequent stages vs. those who did not) were used as the target variables. The number of media exposures were normalised by the number of days when these exposures occurred. In other words, the predictive variables reflected the "density" of media exposures.

The media exposures fell into aforementioned media channels. These channels were planned with different objectives such as awareness, consideration and conversion (**Table 1**). Endemic direct investment, for instance, reflected display advertising on publisher websites that offer content related to home furnishing. Its objective was to build brand awareness among category relevant audiences. The coefficients of the logistic regression models were estimated using the Maximum Likelihood method with the glm() function in R.

Results

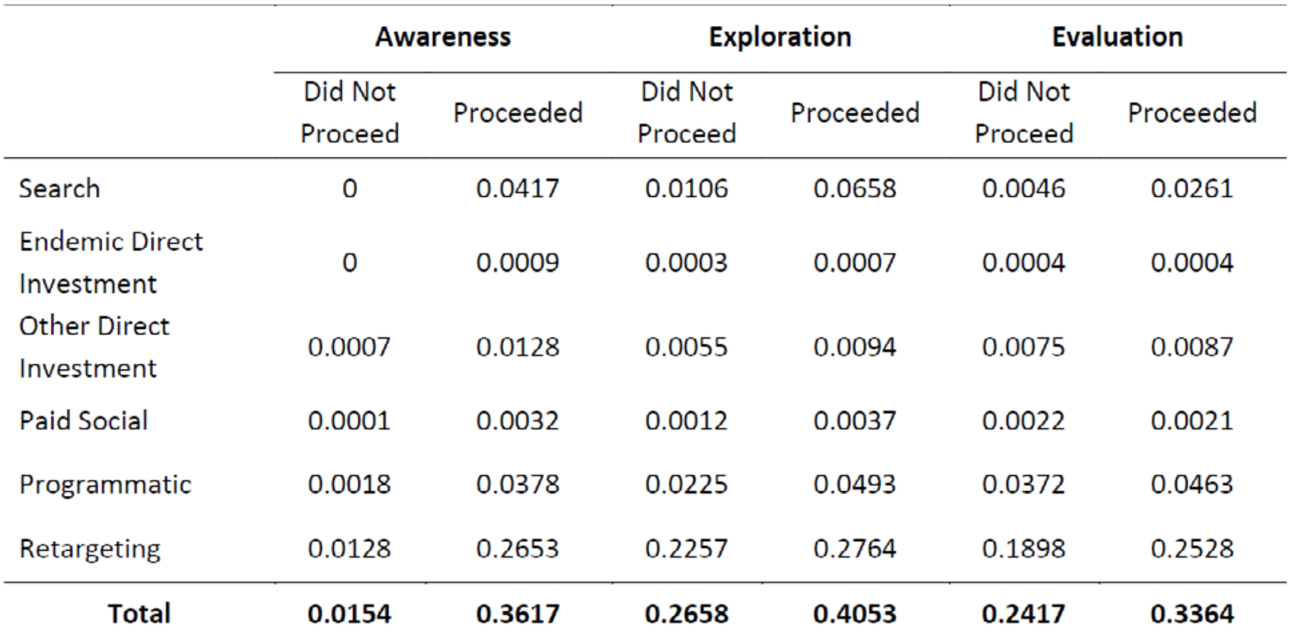
We first obtained a few basic properties of the journey. During the 90-day period, close to a million individuals began the journey with the Awareness stage. Among these, 18% proceeded to the Exploration stage, 7% proceeded to Evaluation stage and finally 3% completed journey with purchases. The results indicated chance of an individual completing the journey at each stage was 3% for Awareness, 17% for Exploration, and 43% for Evaluation.

The average length of a journey was 22 days. The average length of a stage was 3, 9 and 12 days for Awareness, Exploration, and Evaluation respectively. Interestingly, the Awareness stage was relatively short. This may indicate that consumers make a relatively fast decision on whether to further engage with the brand once the first impression was made.

Table 2 shows the cross tabulation of the predictive variables (media exposures per day) and target variables (Proceeded to the next stages vs. who did not) of the logistic regression models. Not surprising, for each stage, the consumers who proceeded received more media exposures than those who did not, indicating that media exposures differentiated these two groups of individuals. For instance, in the Awareness stage, those who proceeded received .362 ads per day whereas those who did not received only .015 ads per day.

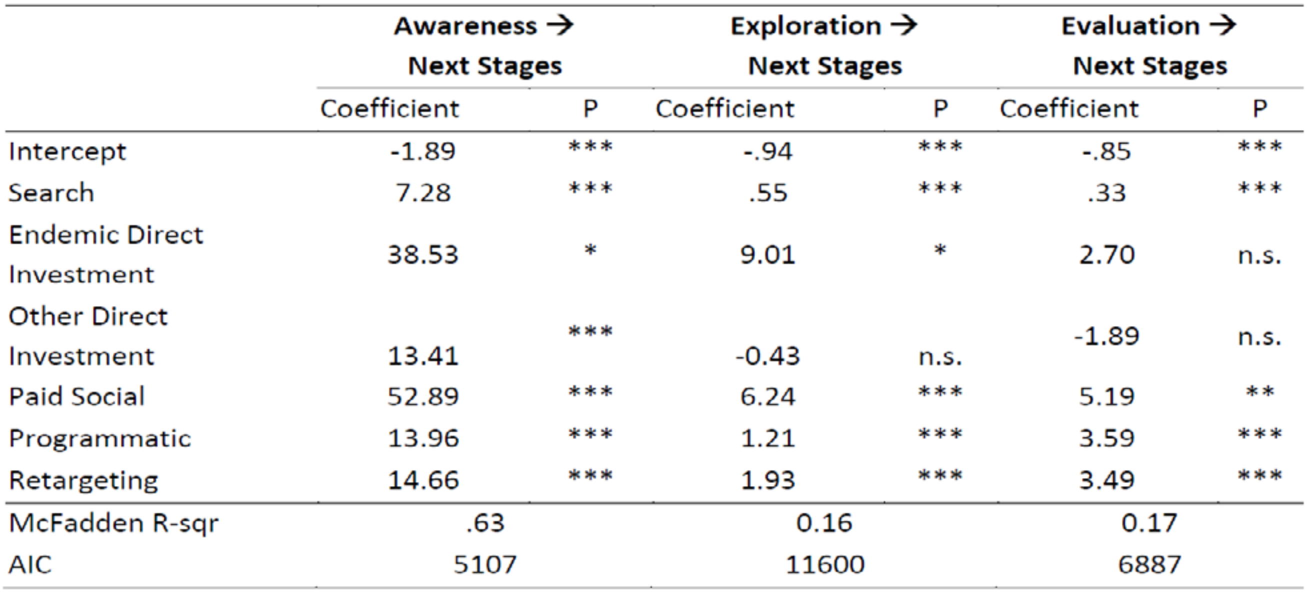
It is particularly interesting that the difference in media exposure levels was highest at the Awareness stages, indicating a greater role of media at this stage. Considering consumers are relatively open to new information at this point, they can be relatively attentive to the brand information in advertisements. After the Exploration stage, the consumers have browsed the brand website quite extensively; there is likely less room for further advertising communication.

Table 3 shows the results of the logistic regression models. Most of the media tactics displayed significant coefficients, indicating they have impacted the progression of the journey. In the Awareness stage, endemic direct investment and paid social showed significantly high coefficients, indicating immense impact of these tactics. At this stage, consumers are relatively far from making a specific product choice; the main task is to understand possibilities. The endemic, home furnishing websites may have provided relevant information that helped consumers explore possibilities. It's not surprising advertisements in such environments is likely to be noticed. Similarly, paid social advertisement



	Awareness		Exploration		Evaluation	
	Did Not Proceed	Proceeded	Did Not Proceed	Proceeded	Did Not Proceed	Proceeded
Search	0	0.0417	0.0106	0.0658	0.0046	0.0261
Endemic Direct Investment	0	0.0009	0.0003	0.0007	0.0004	0.0004
Other Direct Investment	0.0007	0.0128	0.0055	0.0094	0.0075	0.0087
Paid Social	0.0001	0.0032	0.0012	0.0037	0.0022	0.0021
Programmatic	0.0018	0.0378	0.0225	0.0493	0.0372	0.0463
Retargeting	0.0128	0.2653	0.2257	0.2764	0.1898	0.2528
Total	**0.0154**	**0.3617**	**0.2658**	**0.4053**	**0.2417**	**0.3364**

Table 2. Average number of exposures per day by media tactic.



	Awareness → Next Stages		Exploration → Next Stages		Evaluation → Next Stages	
	Coefficient	P	Coefficient	P	Coefficient	P
Intercept	-1.89	***	-.94	***	-.85	***
Search	7.28	***	.55	***	.33	***
Endemic Direct Investment	38.53	*	9.01	*	2.70	n.s.
Other Direct Investment	13.41	***	-0.43	n.s.	-1.89	n.s.
Paid Social	52.89	***	6.24	***	5.19	**
Programmatic	13.96	***	1.21	***	3.59	***
Retargeting	14.66	***	1.93	***	3.49	***
McFadden R-sqr	.63		0.16		0.17	
AIC	5107		11600		6887	

Table 3. Logistic regression results.

may have benefited from targeting audiences who interacted with home furnishing content or posts on social networks.

The impact of endemic direct investment and paid social continued to be strong at the Exploration stage. The task of the consumer becomes exploration of product options towards a more finite consideration set. They have browsed the website quite extensively at this point. Paid social and endemic ads still impacted them, probably because they had not yet finalised the consideration set and were still open to ideas suggested in the advertisements.

At the Evaluation stage, the impact of advertising seemed to have generally decreased as the coefficients of several tactics have either become less or not significant. Endemic direct investment in particular has become non-significant, whereas it was highly effective in the previous two stages. At the Evaluation stage, consumers are likely to have formed a finite consideration set and be more focused on making a decision; the role of advertising messages may have reduced.

It is important to note that the results of logistic regression models presented a different picture of media channels compared with performance reports based on conversion rate, which is shown in **Table 4**. Search and other direct investment were two channels of which the performance appear inflated with conversion rate. The high conversion rate of other direct investment was mainly driven by a geo-fencing tactic and optimisation against conversion rate, both of which resulted in reaching audiences close to purchase. From consumer journey perspective, as we showed earlier, consumers who are close to purchase are pre-disposed to make a purchase in the first place. The high conversion rate did not necessarily indicate high effectiveness of advertising. Our results also showed conversion rate tends to favour search. This is understandable because search reflects a consumer action, which is a result of a consumer's intention or interest in the first place. The incremental benefit of search was quite limited as our results showed.

	Conversion Rate
Search	2.05%
Endemic direct investment	0.64%
Other direct investment	0.57%
Programmatic	0.26%
Retargeting	0.24%
Social	0.49%

Table 4. Logistic regression results.

Discussion

The results of the study have shown differences in advertising impact for stages of the consumer journey. Advertising was highly useful for moving consumers from awareness to deeper engagement with a brand. When a consumer begins a journey, they are open to suggestions; their brand knowledge is limited. The basic information that an advertisement carries; such as the brand name, product images, and call to action messages can be effective in triggering the interest to learn more. At a later stage of the journey, consumers have obtained a significant amount of brand information, so the role of advertising reduces.

This study showed potential problems of commonly adopted media optimisation processes based on simplistic metrics such as conversion rate. Such processes can result in high conversion rates that are disproportionate to the actual behavioural impact media channels make. While it is desirable to target audiences close to purchase and demonstrate the results with a high conversion rate, marketers should be aware that consumers are likely to be least receptive to advertising when they get close to purchase. Marketers should not mistakenly interpret high conversion rate as high effectiveness, particularly when making fund allocation decisions.

We acknowledge the performance results of the current study reflected only a single case and cannot be generalised. Advertising strategies can vary tremendously by campaign, product, and category. Our intention is not to draw general conclusions regarding the relative performance media channels. Individual studies are required to truly understand the performance of specific campaigns.

In summary, this study presents an innovative approach to the attribution problem. This approach gives advertisers insights into not only which channel impacts consumer behaviour but where the impact is made. It provides a practical framework for evaluating upper and lower funnel media channels. We believe it adds to the body of knowledge for advertising effectiveness measurement.

Reference

1. D. Edelman, M. Singer, Competing on Consumer Journeys, November 2015, Harvard Business Review.

2. D. Court, D Elzinga, S Mulder, O. J. Vetvik, The Consumer Decision Journey 2009, McKinsey Quarterly.

3. Kotler, P., Keller, K. L. (2012), Marketing Management, Fourteenth Edition, Pearson Education.

Authors

Shawn Song is a Director of Marketing Sciences at PHD Media. Shawn leads media measurement, insights, and analytics for a few PHD clients. Shawn developed and improved several proprietary analytical processes for the company. His work has been recognised by important industry organisations. Prior to PHD, Shawn was an Associate Research Director at the Advertising Research Foundation, where he managed research projects and publications. Shawn received a Ph.D. degree in Textiles and Clothing at Iowa State University. He has also published multiple papers in peer-reviewed journals and professional conferences.

Charlotte Ma was a Senior Analyst of Marketing Science at PHD media, USA at the moment of the writing of this article. Currently she is a Senior Associate for Data Science & Analytics at SapientRazorfish. Her expertise includes consumer segmentation, business analytics, business intelligence, and data visualisation. Prior to PHD Media, she conducted quantitative analysis and survey design at Nielsen, Cache, and Tracx. She has a strong interest in consumer insights research and assessment of advertising media impact on consumer behavior. Charlotte Ma has a Master's degree in Integrated Marketing at New York University, USA.

Pranav Patil was part of the Marketing Science team at PHD Media, USA at the moment of the writing of this article. Currently he is a Senior Analyst at Annalect. His work is focused on examining performance of different media channels and tactics on business results. He began his career in sales and developed a passion for Data Analytics. He devotes his professional time towards furthering knowledge and capabilities in Marketing Analytics. Pranav has a MBA degree in Marketing Management at Pace University, USA and a BBA in International Marketing at Kingston University, UK. He is currently pursuing a Master's degree in Analytics at Harrisburg University of Science and Technology, USA.

Appendix

Description of logistic regression

Logistic regression is a commonly used method to model binary outcomes, that is the target variable y takes the value 0 or 1. It does so by modelling the probability of the target variable y equals 1. It takes the below mathematical form

$$\Pr(Y=1) = Logit^{-1}(X),$$
$$X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots,$$
$$Logit^{-1}(X) = \frac{e^X}{1+e^X},$$

where $\Pr(Y=1)$ is the probability of the target variable y takes the value 1, $X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$ is a linear combination of predictive variables, and $\mathrm{Logit}^{-1}()$ is a function that transforms the continuous value of the linear combination into a probability value between 0 and1.

The parameters of logistic regression β_0 and β_i can be estimated using the Maximum Likelihood method. The log likelihood function takes the form

$$\sum_1^n (log\ (Logit^{-1}(X_i)\ Y_i + 1 - Logit^{-1}(X_i)\ (1 - Y_i))),$$

where i indicates the observations in the data and n is the total number of observations. Estimation of the parameters involves an algorithm that maximises the value of the log likelihood function. Many statistical packages offer maximum likelihood estimates for logistic regression. The current research utilised the glm() function in R.

A Comparison of Machine Learning Approaches for Cross-Device Attribution

Robert Stratton
Neustar

Dirk Beyer
Neustar

Classifications, Key Words:

- Cross-device
- Attribution
- Machine learning

Abstract

We conducted a study to develop practical guidelines for choosing feature generation approaches and classifiers for the problem of cross-device entity resolution. The study was conducted on a synthetic data set, generated by extracting event logs from an agent based geographic simulation of daily mobility patterns in a city. We tested a combination of deep and shallow classifiers against two different feature representations, and also evaluated the impact of observation sampling on classification accuracy. We found that Random Forest models developed on hand crafted features performed equally well to multilayer perceptrons working on untransformed data, but with a lower computational cost.

Introduction

Many applications of data science are concerned with finding associations between inputs and outputs, answering questions about how a particular variable, X is associated with another variable Y. For example, in a digital attribution analysis, X might represent some form of online advertising, and Y a purchase of a product. Increasingly, as more events are logged, and more granular data becomes available, we want to study the relationships between the inputs and outputs at the most granular possible level. Questions that might previously have been addressed using aggregated data, e.g. looking at variation over time or across large cross-sections like geographies, are increasingly amenable to much more granular, entity level analysis.

But before we can identify any relationships at a granular entity level, we need to be sure that we can connect the relevant inputs to the relevant outputs – to be sure that the person who saw an ad was the same person that made a purchase, even if the two events happened on different platforms. Although digital event tracking and cheaper data storage has made more granular analysis possible, the proliferation of digital devices and services means that many events are tracked on different systems and as users move across these devices and services to complete their tasks, their identity and histories becomes fragmented. Compounding this is the fact that the data associated with a specific identity fragment is itself increasingly diverse and unstructured. An identity fragment may be described by a variety of information, all referring to the same individual, but scattered across, images, text, JSON and XML, obscured by noise of various kinds. The science of making these connections – entity resolution – is an important pre-requisite for the success of many data science projects, and in digital attribution, is critical to assigning credit to different marketing channels in a way that correctly measures the value of each marketing contact.

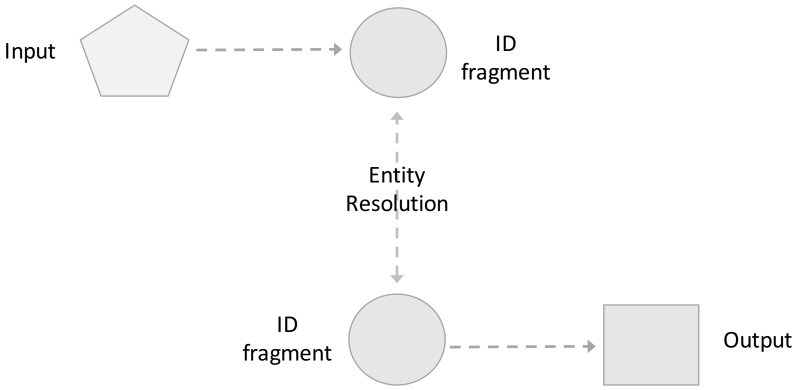


Figure 1: Resolving ID fragments to create an association between inputs and outputs, connected through an integrated identity

Figure 1 shows an example of two fragmented identities that need to be re-constituted in order for a relationship between input and outputs to be discovered.

Similar challenges related to associating elements arise in other fields, such as social network detection (Faloutsos 2004), document and image collection management (Kyle Heath 2010) (Hossain 2012), but the sheer volume, as well as the social and commercial significance of data related to human attributes and actions makes it a key area of research for entity resolution.

In this paper we'll review some of the main factors that can determine the success or failure of probabilistic cross-device resolution in the context of human behaviour. More specifically we will look at how each of the following elements can impact resolution of digital devices back to an underlying user:

- Data transformation and feature generation approaches
- Classification algorithm selection
- Interactions between data transformations and classification algorithms

Methodology Overview

To evaluate the importance of each of the factors outlined in the previous section, we needed an experimental framework in which we could test different methodologies and their combinations, as well as benchmark their predictions against a known, labelled outcome. The use of real datasets presented a tantalizing proposition, but these can present problems in obtaining an accurately labelled truth set against which to benchmark methodologies. In addition, real datasets can be problematic due to privacy concerns, so we were ultimately limited to data that was either synthetic, or already in the public domain. We looked at a number of mobility dataset sources, including the GeoLife Trajectory, the MIT Reality Commons, and the LifeMap datasets, but due to sample sizes and other characteristics of the data available, we opted to use a simulated dataset instead. Many similar simulations of this type have already been published, including several by Harland and Birkin (Birkin 2013) (Harland 2013). In addition to its open-source characteristics, simulation offers several other important advantages over real data. Firstly, we know the real answers, so we can benchmark the accuracy of different methodologies against the actual data generating process. Secondly, we can examine the impact of counter-factual scenarios, allowing us to look at what would have happened under different conditions.

Simulation

The framework we developed is an agent-based geographic simulation, based on elements from existing published studies, which provide the basis for agent's navigation strategies and their destinations, with innovations added to control agent mobility cycles and device usage. The simulation is deliberately abstract in that it doesn't aim to capture all aspects of human behaviour but rather the main elements of daily mobility – cycles of movements within a day – for a collection of individuals in a simulated city environment. The ability of relatively simple processes to represent the main features of these mobility patterns has been proven by others including Eagle, who argues that although humans can potentially display relatively random patterns of behaviour, there are easily identifiable routines in every person's life. (N. a. Eagle, 2006).

In our simulated system, each agent represents a human in the geographic space. We use two elements from a NetLogo simulation, which accompanies the paper 'Agent-based simulation of human movement shaped by the underlying street structure' by Jiang and Jia (Jiang, 2011). These elements are an urban geography, extracted from an area of North London, and an agent street navigation system.

The original intent of the Jiang and Jia paper was to look at how vehicle traffic and pedestrian flow is impacted by the underlying street structure, and we added three types of activity containers to the geography, connected by the original street structure. The activity containers were:

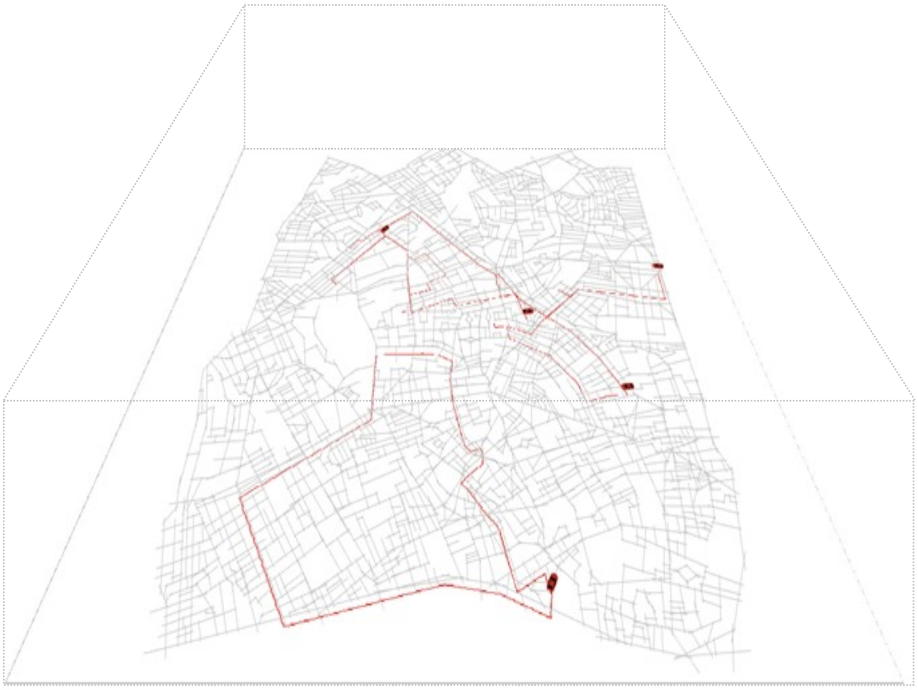


Figure 2: Sample paths taken by agents around the street system as part of their daily mobility routines

1. **Houses** – distributed around the map
2. **Workplaces** – clustered in the centre
3. **Public spaces** – distributed around the central and outer areas

We gave each agent a home, workplace and public space preference, and a routine based on principles from Eagle and Pentland (N. a. Eagle, 2009). The routine essentially sets approximate times at which an agent's goal changes to be at a location which is different to their current one. Each agent has a slightly randomised routine, offset by up to 3 hours from the mean routine, and each travel event commencement is also slightly randomised from the offset.

We also gave each agent between 1 and 3 different digital devices, and the ability to 'use' the devices. Use of the device, for our purposes, means the creation of an event log that records aspects of the usage event. With each device usage an event is logged which gives the device identifier, a date time stamp and a tile identifier. Each tile is intended to represent a location-identifier, which in the real world could be a GPS location, WiFi network or IP address. We exclude other device features that may be available in real world situations, such as browser type, operating system, make and model. The tile identifier is based on a partitioning of the street system into 2401 areas of equal size – see **Figure 3.**

A simple in-built cycle controls use of these devices – every 120 seconds the agent uses one of the devices he holds. These different digital device usage logs represent the fragmented ways in which the identity of the individual can be observed, which we seek to resolve back to the underlying agent through the entity resolution process. The simulation was run with a 12 second time granularity – in other words every iteration represents a 12 second period of elapsed time. It was run over 72,000 iterations, equivalent to 10 simulated days.

Methodology Setup

Figure 4 shows the overall flow of the analytical process we use in this section. The simulation outputs the device usage logs in the form:

[device-id, date-time stamp, tile location]

Given these logs, we want to learn a classifier that allows us to allocate device-ids into groups that reflect the device pools used by the underlying agents. In other words, we want to be able to assign each device to a group, such that all the devices in that group belong to the same agent. The key information available is regularity in the agent's movements, their persistence at specific locations between travel activities.

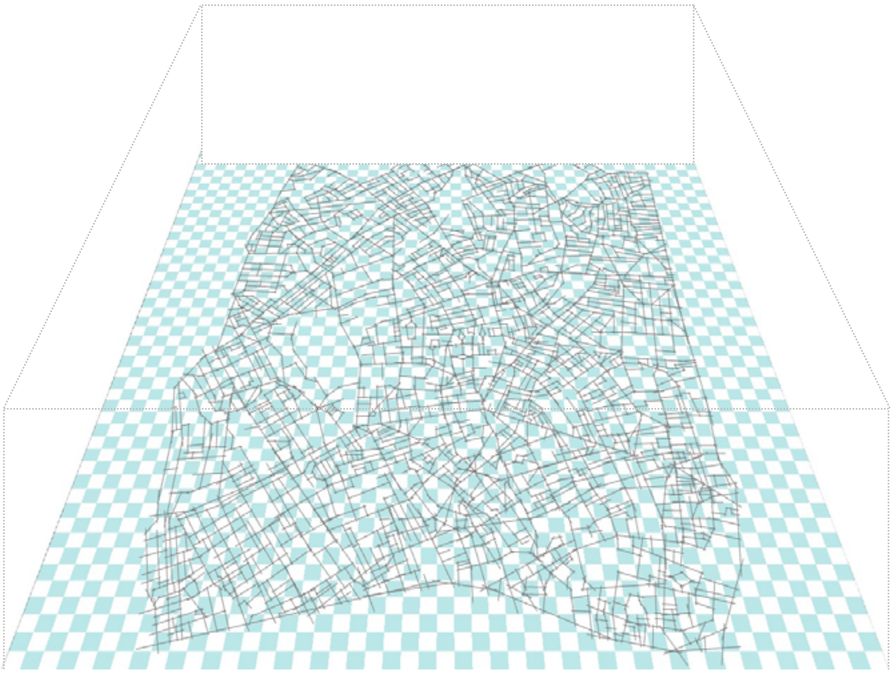


Figure 3: The 2401 tiles that partition the simulated geography

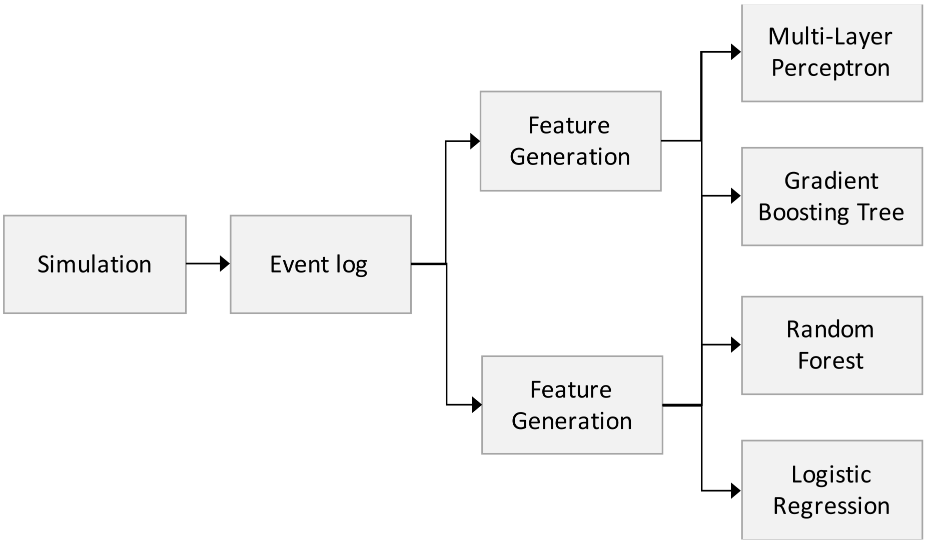


Figure 4: Overall analytical process

In this section we consider several elements that may have an impact on our ability to re-connect the devices into clusters.

1) The role of feature generation – how to represent the input to the classifier, what is the best representation of the sample data to learn a solution to the problem.

2) The importance of the role of classifiers, in terms of their

accuracy and computational complexity. This set is called the hypothesis space of the learner. If a classifier is not in the hypothesis space, it cannot be learned.

Feature Generation

In order to understand whether the entity resolution is best tackled using a shallow classifier and a hand-engineered feature space, or a deeper classifier and an automatically learned representation, or some other combination, we looked at two main feature representations, one that involved pre-coding hypotheses as features, and the other a simple count of overlaps per tile.

Manual Feature Generation

The events for each device ID could be thought of as a bag of words, a list of tuples reflecting the tiles on which the device has been observed and the number of times it has been observed there, for example:

device-id 7: *[(tile-1, 10),(tile-2390, 2)].*

device-id 12: *[(tile-55, 3),(tile-2390, 5),(tile-1753, 6)].*

We used the intuition that the basis of the signal that could connect entities in this system was commonality in the tiles observed for each device, and the frequency of visits that each tile was associated with. But it also seemed likely that some tiles would provide less distinctive signals since they were likely to be visited by a larger number of agents as part of their daily cycle.

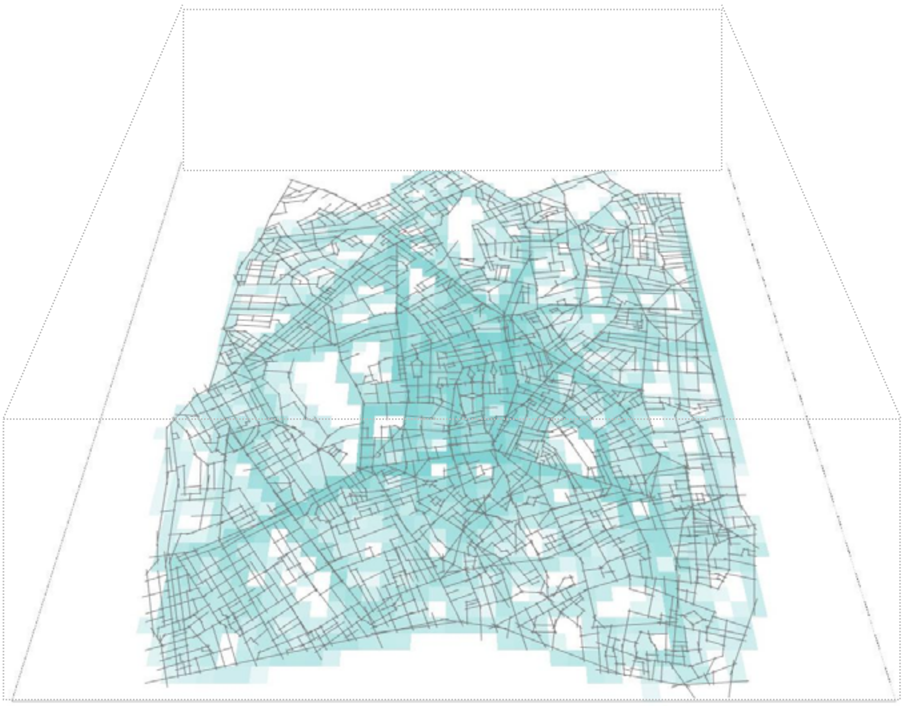


Figure 5: The count of unique devices ever observed on each tile, darker shades imply higher observations

We calculated the number of unique agents ever observed per tile and plotted it back onto the map, showing a concentration of tiles around the centre of the street system. This appeared to be due to two main factors. Firstly, as observed by Jiang and Jia (Jiang 2011), agents are attracted to the straighter, more direct routes crisscrossing the centre of the map in navigating from point to point and are therefore more likely to have been observed on these streets. Secondly, the workplace and public space containers are in the central part of the map, and so agents are more likely to have visited these areas as part of their daily mobility patterns.

It seemed likely that observing a co-occurrence of devices (i.e. the minimum of the counts of visits to the tile by any pair of devices) on some tiles will be less useful in identifying true pairs, since many other entities share the same tile as part of their histories. To control for the level of activity on each tile, we calculated an Inverse Document Frequency transformation for each tile.

$$idf(t, D) = log \frac{N}{1 + |\{d \in D : t \in d\}|}$$

Where N is the total number of device histories, and $\{d\in D{:}t\in d\}$ is the number of device histories where a tile appears.

Plotting the IDF score back onto the map by tile (**Figure 6**) we can see that the central system has been down-weighted.

The final feature set for this representation contained 3 features for each candidate pair of users:

- The total count of events on which each device-ID was observed.
- The total number of tile visit events in common between each history.
- The total number of tile visit events in common divided by the idf term essentially applying a penalty to overlap events on tiles which are generally more likely to be visited.

Image Representation

The second feature representation we used was much simpler and involved less pre-processing. For each pair of device-ids observed we calculated the number of occasions they had visited each tile, then took the minimum count between the pair, per tile. The data therefore has one variable for each tile. Where there is no common visit to a tile the cell value will be zero. Plotting the counts back onto the map for a pair of device-ids that do belong to a common agent, we can see a pattern that could reflect a daily mobility cycle for one agent.

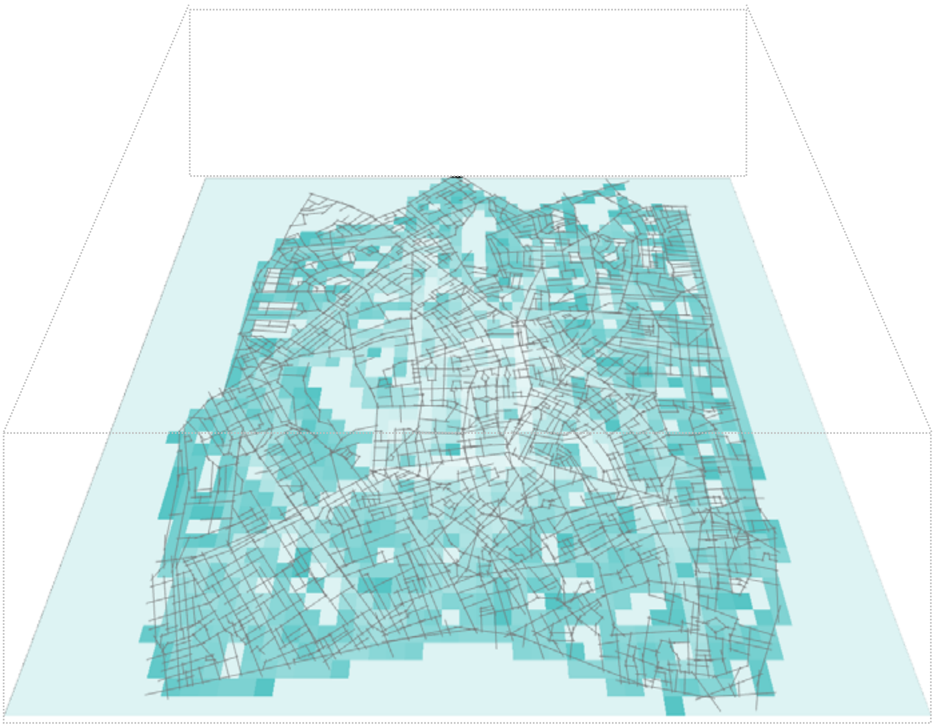


Figure 6: The IDF score for each of the tiles in the simulated data. Darker shades imply a higher score

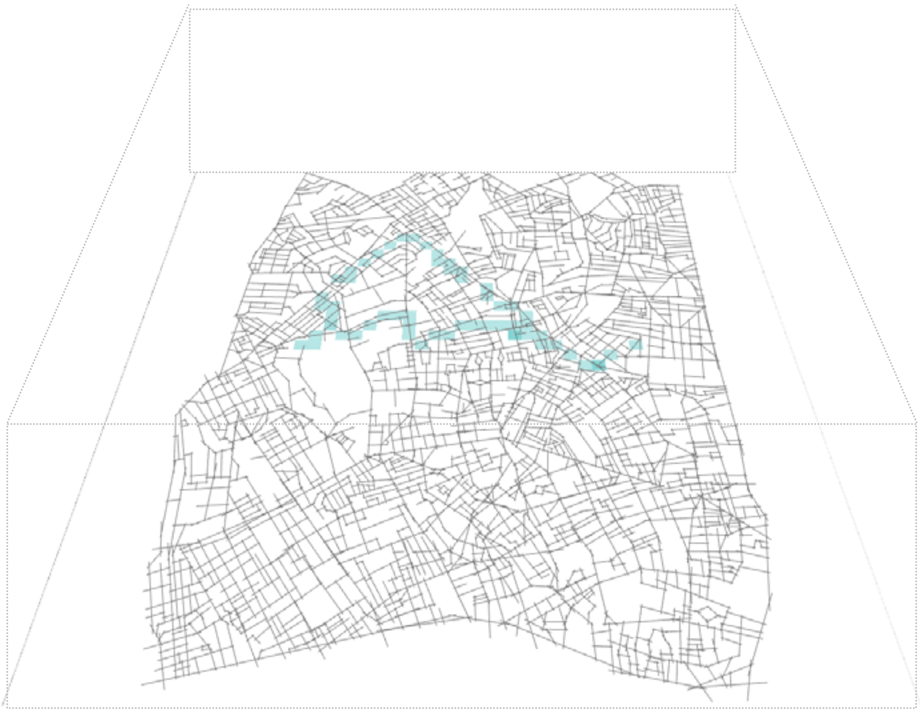


Figure 7: An example of one of the images reflecting pairwise tile overlap presented to the classifiers

With this representation we are not providing any hypothesis about which tiles are more likely to be useful in identifying commonality, but expecting that the classifiers will learn which tiles to assign more or less importance to.

Classifiers

Several studies have already been conducted assessing classifier performance (Fernández-Delgado, 2014) (R. N. Caruana, 2008), (R. a.-M. Caruana, 2006) but none to our knowledge on an entity resolution problem. Like the feature representation, the choice of classifiers is important since it defines the hypothesis space that can be tested.

We tested the following classifiers:

- **Random Forest**, 45 estimators and a max. tree depth of 30, using Spark ML
- **Logistic Regression**, solved with LBFGS, using Spark ML
- **Multilayer Perceptron**, layers (10,10,5,5), using Tensorflow
- **Gradient Boosting Tree**, 45 estimators and a max. tree depth of 30, using Spark ML

Once we had developed the full feature stack we split the sampled dataset into test and train samples, with 70% of the data assigned to the training set and 30% assigned to the test set. N-fold testing was impossible due to the number of sampling steps and the already expensive computational requirements. We evaluated the models using Recall and Precision, typically used in information retrieval, and their harmonic mean, the F1 score. We preferred to use these scores since they are independent of the sampling rate, and the data was already highly skewed (negative examples/ positive examples) due to the large numbers of non-matching cases in the pairwise representation.

Both the Multilayer Perceptron and Logistic Regression classifiers improve their accuracy substantially on the higher dimensional image dataset, while the two tree-based models see major declines. The top 3 performers are very close, with less than a 0.2% between them, but the classifiers perform very differently across data representations, suggesting interaction between the classifier and the data representation. The most consistent performance across the two representations is from the Logistic Regression model, which has less than a 6% comparative difference.

Plotting the logistic regression coefficients from the image representation back onto the map (**Figure 10**), we can see that the busy areas around the centre of the map are given lower coefficients, indicating that to some extent it has

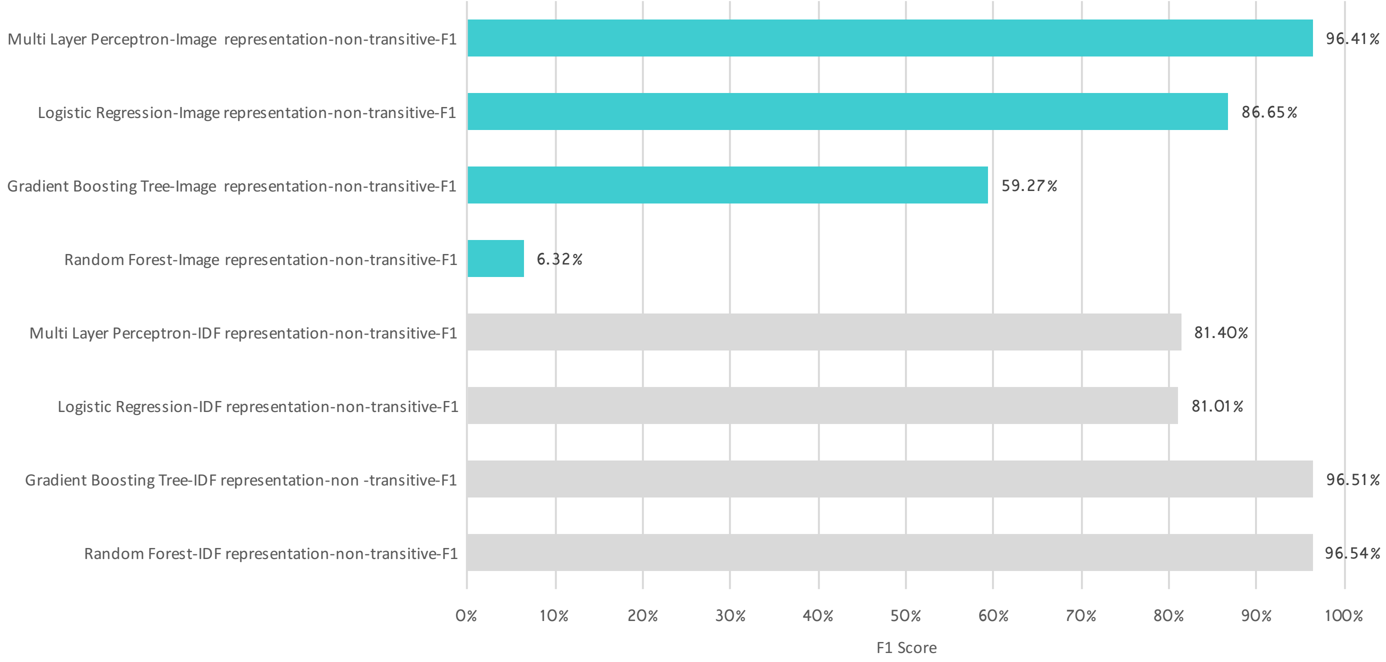


Figure 8: Results for the 2 feature representations and 4 classification algorithms used

detected areas that deserve less weight, similar to the IDF score. But its low performance relative to the highest scoring classifiers suggests that the problem is more non-linear than can be tackled with an algorithm that is designed to separate cases along a hyper-plane frontier. The relatively high score of the Multilayer Perceptron compared to the Logistic regression on the image dataset seems to confirm that there are non-linearities that can be handled by the introduction of hidden layers.

The results we see are broadly consistent with findings from other studies, with a few exceptions. Domingos argues that many algorithms that work well in low dimensions become intractable when the input is high-dimensional. (Domingos, A few useful things to know about machine learning, 2012). This conforms with what we see in the tree-based model's performance. In addition, the results conform with the experiments in (Caruana & Niculescu-Mizil, 2006) where boosted decision trees are seen to perform well across a range of problems when dimensionality is low. The better performance of the Multilayer Perceptron on the higher dimensional dataset is consistent with the suggestion by LeCun et al (LeCun, 2015) that feature generation is unnecessary if the features can be learned automatically by a general-purpose learning procedure like a deep neural network. The superior performance of the Logistic Regression Classifier on the highly dimensional dataset is in conflict with expectations that in general 'shallow' classifiers need pre-processing to extract features that can select the relevant parts of the data.

Computational Cost and Scalability

Figure 11 shows the mean CPU times for each of the classifiers on the two different data representations. Although the F1 score for the Multilayer Perceptron on the image representation is close to the Random Forest on the IDF, the run-time is significantly longer. In addition, the Multilayer Perceptron is unlikely to scale well to an increased number of tiles, since each new tile requires as many new coefficients as are in the first hidden layer.

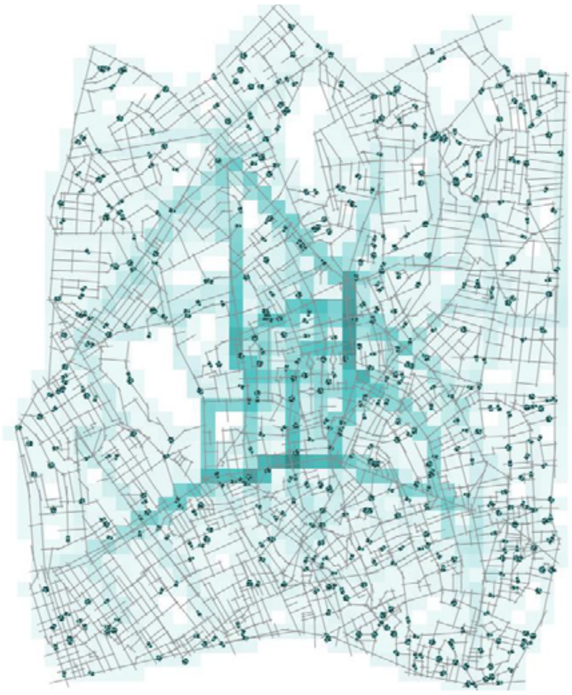


Figure 9: Count of unique devices ever observed on each tile

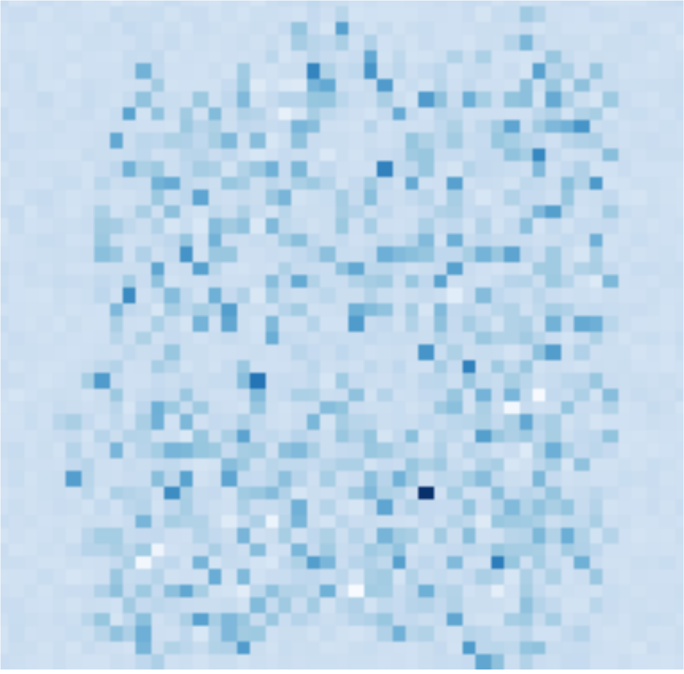


Figure 10: Coefficients from the logistic regression model plotted on a 49*49 grid

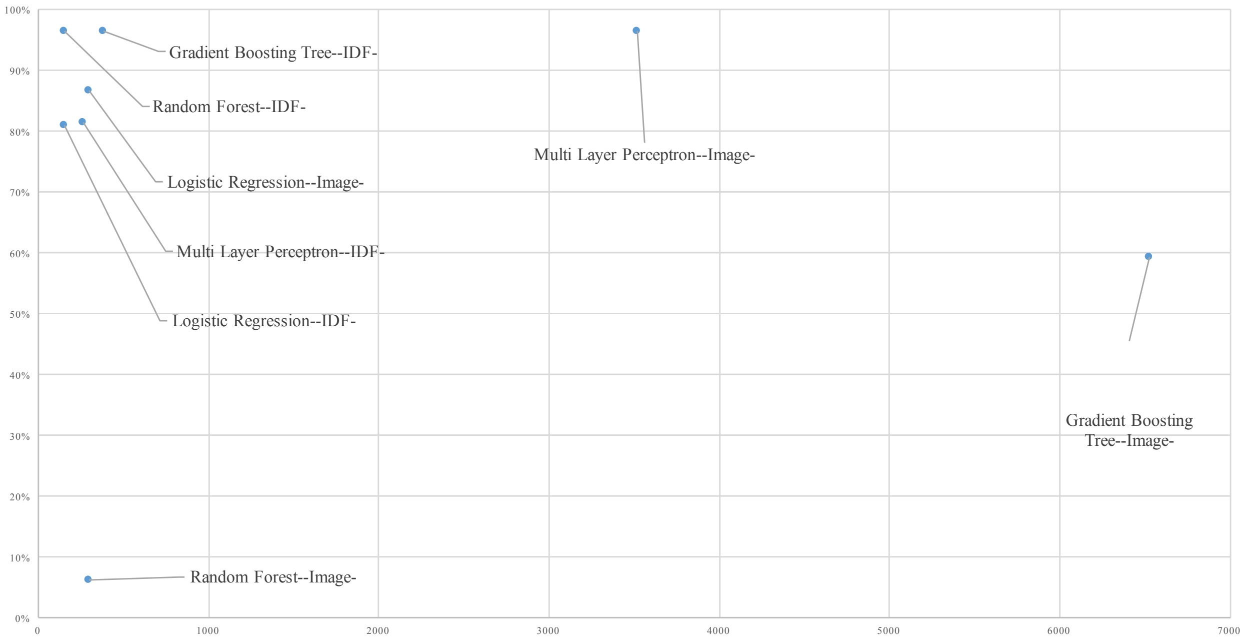


Figure 11: Average computation speed vs. F1 accuracy of different algorithms on the different data representations

Conclusions

Overall, we observed significant variability in accuracy across data representations and classifiers, as well as a wide range in the performance of different classifiers across the data representations. Given that the three top performers spanned different representations and were almost tied in terms of their accuracy, other considerations such as feature development time and computational complexity should also be considered. If the same modelling exercise is to be performed repeatedly, the higher time investment in hand-crafted feature development would pay off in return for modelling CPU costs. However, for a one-time exercise, the MLP performed well, with limited feature development time and a model calibration time that was still in minutes rather than hours. The scalability of MLP is likely to be a problem though, as additional tiles add multiple new coefficients that each need to be estimated.

References

1. Birkin, Mark and Harland, Kirk. 2013. "Simulating retail demand at the individual level: stage 1 demand synthesis."
2. Brunsdon, Chris,, and Alex Singleton. 2015. Geocomputation: a practical primer,. SAGE Publications.
3. Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An empirical comparison of supervised learning algorithms." Proceedings of the 23rd international conference on Machine learning.
4. Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina. 2008. "An empirical evaluation of supervised learning in high dimensions." Proceedings of the 25th international conference on Machine learning.
5. Domingos, Pedro. 2012. "A few useful things to know about machine learning." Communications of the ACM.
6. —. 2015. The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

7. Eagle, Nathan, and Alex Sandy Pentland. 2006. "Reality mining: sensing complex social systems." Personal and ubiquitous computing.

8. Eagle, Nathan, and Alex Sandy Pentland. 2009. "Eigenbehaviors: Identifying structure in routine." Behavioral Ecology and Sociobiology.

9. Easley, David, and Jon Kleinberg. 2010. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press.

10. Faloutsos, C., McCurley, K.S., Tomkins, A. 2004. "Fast discovery of connection subgraphs." Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

11. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. 2014. "Do we need hundreds of classifiers to solve real world classification problems." J. Mach. Learn.

12. Harland, K., and M. Birkin. 2013. "Using Synthetically Generated Populations in Agent-Based Models."

13. Hossain, M.S., Gresock, J., Edmonds, Y., Helm, R., Potts, M., Ramakrishnan, N. 2012. "Connecting the Dots between PubMed Abstracts." 10.1371/journal.pone.0029509.

14. Jennings, Michael Wooldridge and NR. 1995. "Intelligent agents: Theory and practice." The Knowledge Engineering Review.

15. Jiang, Bin, and Tao Jia. 2011. "Agent-based simulation of human movement shaped by the underlying street structure." International Journal of Geographical Information Science.

16. Kimmig, A., Bach, S., Broecheler, M., Huang, B., & Getoor, L. 2012. "A short introduction to probabilistic soft logic." Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications.

17. Kyle Heath, Natasha Gelfand, Maks Ovsjanikov1, Mridul Aanjaneya, Leonidas). 2010. "ImageWebs: Computing and Exploiting Connectivity in Image Collections." Computer Vision and Pattern Recognition (CVPR).

18. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." Nature.

19. Luger, George F. 2005. Artificial Intelligence: Structures and Strategies for Complex Problem Solving. Pearson Education.

20. Neslin, S.A, and V Shankar. 2009. "Key issues in multichannel customer management: current knowledge and future directions." Journal of Interactive Marketing 23 (1): 70-81.

21. Richardson, Matthew, and Pedro Domingos. 2006. "Markov logic networks." Machine Learning.

22. Sondhi, Parikshit. 2009. "Feature construction methods: a survey." Univeristy of Illinois at Urbana Champaign.

Authors

Robert Stratton, PhD, is a Senior Group Director at Neustar, where he works on R&D projects related to digital identity, multi-touch attribution and panel based models. He has a PhD in Informatics and has worked in the field of marketing analytics for 20 years, previously at WPP.

Dirk Beyer, PhD, is the Head of Data Science Research at Neustar, Inc., where, among other things, he leads the company's marketing measurement innovation and identity analytics. He has held senior analytics science roles at various companies including MarketShare, M-Factor (acquired by DemandTec), and Hewlett-Packard Laboratories.

Using Ensembles to Solve Difficult Problems

Steven Struhl
Converge Analytic

Classifications, Key Words:

- Ensembles
- Machine learning
- Classification trees
- Marketing science
- Prediction

Abstract

Ensembles are not that well known, but worth learning and applying. They can solve thorny or otherwise intractable problems. They also can outperform other methods in predictive accuracy. We will examine the use of several ensemble methods: for instance, showing the strengths of effects of many types of variables and predicting outcomes with high accuracy. We will discuss the underlying workings of these methods, their best applications and limitations. Several real-world examples will be included. This article will explain these methods and their practical applications clearly and will not require that the reader has complex statistical knowledge.

Using ensembles to solve difficult problems

Ensembles, aside from the random forests methods favoured by marketing scientists, are not that well known. But they are worth learning and applying. They also can outperform other methods in predictive accuracy. We will examine the use of several ensemble methods in, for instance, showing the strengths of effects of many types of variables, and in predicting outcomes with high accuracy.

Ensembles combine the estimates of many models by either averaging or voting. They capitalise on one of the key discoveries from machine learning. The average of many indifferent or weak models typically is better than any of the individual models. Ensemble methods usually run dozens or hundreds of models and get a consensus from them.

Ensembles have one disconcerting property: while they may perform amazing feats, they can tell us little or nothing intelligible about what they actually are doing. We must take the output and trust that the means to reach it are valid.

As mentioned, the most known ensemble method is random forests. It is based on classification trees. Even if classification trees are a somewhat hazy concept to you, the word "forest" suggests that many trees are involved. You will find anywhere from 500 to 1000 classification trees taking part in a typical analysis.

Let's review classification trees first, in case these are not a part of your everyday life. This method splits and re-splits a sample to get small groups that differ strongly in terms of some target variable. The target variable could be amount of money spent, how many people fall into specific groups, or—as in our small example below—what percentage consumes everybody's favourite breakfast-like substance, SoggyOs.

In this example, about 50 possible predictor variables were tried. These included a range of household characteristics, such as income, number of adults at home, type of residence, ages of adults and children, and the ones appearing in **Figure 1**, town size/type, and number of children at home.

The procedure scanned through all possible ways in which to split the sample using these predictors, starting with the total sample of 1455, appearing in box (or **node**) 1. The best predictor, leading to the strongest contrast, was town **type/size**. The procedure found the optimal contrast by putting those living either in city or rural areas in one group (in node 6), and those living in the suburbs in another group (in node 2). The difference is modest, with 21.9% eating this substance in the suburbs, and 17.2% eating them elsewhere.

Things get more interesting once we look within the suburban group only. This type of splitting and re-splitting of a sample is what makes classification trees so powerful. Here we see that number of children at home emerged as leading to the sharpest contrasts. The procedure divided this variable into three groups: one or two children (in node 3), 3 children (node 4) and 4 or more (node 5). Everybody in this sample had to have children, and they had up to an astonishing 15 living at home.

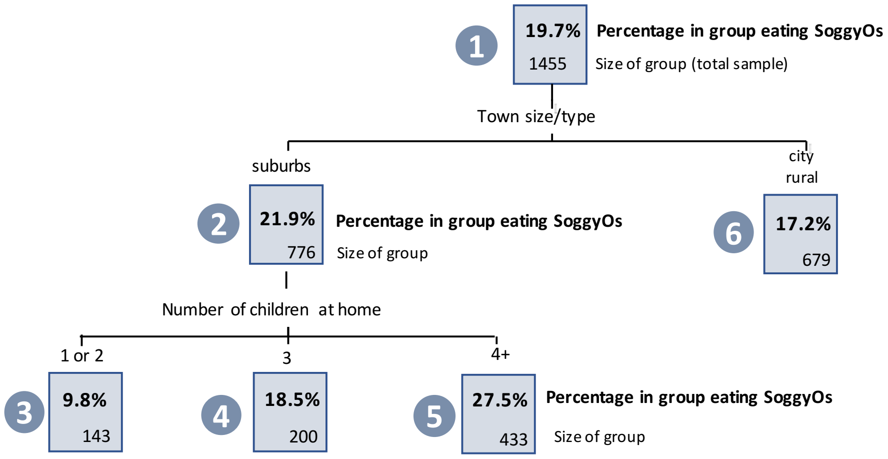

1
19.7%
1455
Percentage in group eating SoggyOs
Size of group (total sample)
Town size/type
suburbs
city
rural
2
21.9%
776
Percentage in group eating SoggyOs
Size of group
6
17.2%
679
Number of children at home
1 or 2
3
4+
3
9.8%
143
4
18.5%
200
5
27.5%
433
Percentage in group eating SoggyOs
Size of group

Figure 1: A section of a classification tree

In this three-way split, we see that one group (living in the suburbs with 4+ children, in node 5) is nearly three times as likely to consume SoggyOs as another (living in the suburbs, but with one or two children, in node 3). The respective levels are 27.5% and 9.8%.

The classification tree would continue splitting the sample into smaller groups until it reached a pre-set minimum size, or it ran out of predictor variables that produced statistically significant differences between groups split off at any given spot. The result would be a set of small groups with very high incidences of consuming SoggyOs, a set with very low incidences, and a set with about average incidences.

These groups would be fully described in terms of their demographic characteristics. These descriptions involve no equations, but rather simple "if-then" rules. For instance, the rule for node 3 in diagram 1 would run like this: If type of town = suburbs AND number of kids at home = 1 or 2 THEN incidence of consuming SoggyOs = 9.8%.

Also, everybody would belong to one and only one group. And groups would include everybody. That last characteristic is sometimes described by the more prolix phrase, "the groups are exhaustive and mutually exclusive."

Our First Ensemble: Random Forests

While you get an easy-to-apply model with classification trees, you may have some other questions. These models are highly dependent on which variable gets entered in a top. Since this method does not look forward, this first choice could be a bad one overall for the total model. You might do better overall if you chose another variable that was nearly as good at that spot, but which allows better splitting below.

Therefore, it is logical to ask, which variables are truly important? Also, is there some way to make the model sturdier overall?

As you might suspect, we find the answers to both questions in random forests. This method runs many hundreds of classification trees, in each tree swapping out people and variables at random. Then all the trees "vote" on an outcome.

An excellent reading of true variable importance comes from observing how classification levels change as variables and people are moved into and out of many models. The analysis also provides diagnostics showing how variables' importance shift as more trees are run and their estimates are added. Eventually, after running a few hundred alternative models, we reach stable estimates.

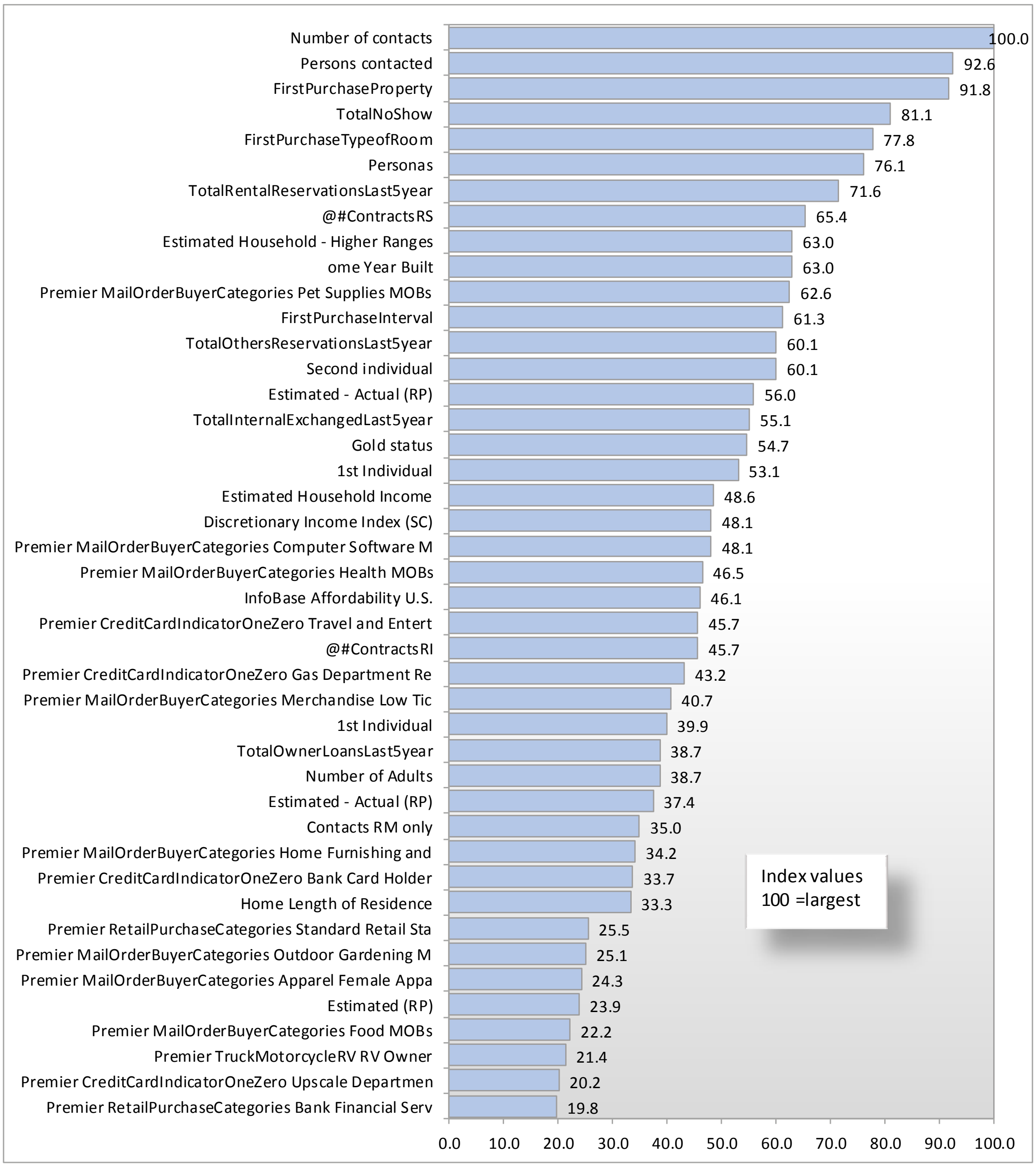


Figure 2: Some importance scores from random forests

This procedure allows us to get both an overall correct prediction score and a fix on variables' importance. An example of some variables' importance appears in **figure 2**. This reflects only the top 45 out of over 200 analysed in one study. The relative effects are clear. However, this model is an average across several hundred trees where variables and people were swapped in and out randomly in each run. We could not look across this mass of tree models and see any structure that we would even faintly understand.

Yet an individual tree model could be far from optimal, if it happened upon a bad starting variable. There is a way to try to get the best of both worlds, or at least some of both. We can run random forests first among all possible variables, and find those that have the strongest effects across the many hundreds of trees involved. Then we can build one final classification tree model using just the variables emerging as strongest. This one model is clear and easy to understand, and based on "assured winners." We can then use this with confidence as a sound guide to decision making.

By the way, "random forests" adds to the list of horrible names beloved of math and science types (think of SCSI –pronounced "scuzzy"— drives, box and whisker plots, and p/p plots, for instance). Still, random forests illustrate a key finding from machine learning. That is, the average of many weak estimates typically is better than any of the individual estimates. Each run of an ensemble takes a slightly different view of the data. It is only by taking these different views that ensembles gain their great value.

Finding more than one value of a predictor as important: Boosted decision stumps

Any approach that uses many models, getting an average of estimates, is called an **ensemble** method. With ensembles, we have ventured deep into **machine learning.**

The next ensemble application gives us another, distinctive way to determine variables' importance. It too has a fairly awful name: **boosted decision stumps**. It uses a process of building single-level trees repeatedly. Its approach differs from that of random forests, which reruns larger trees while randomly swapping predictor variables and cases (people) in and out of each model.

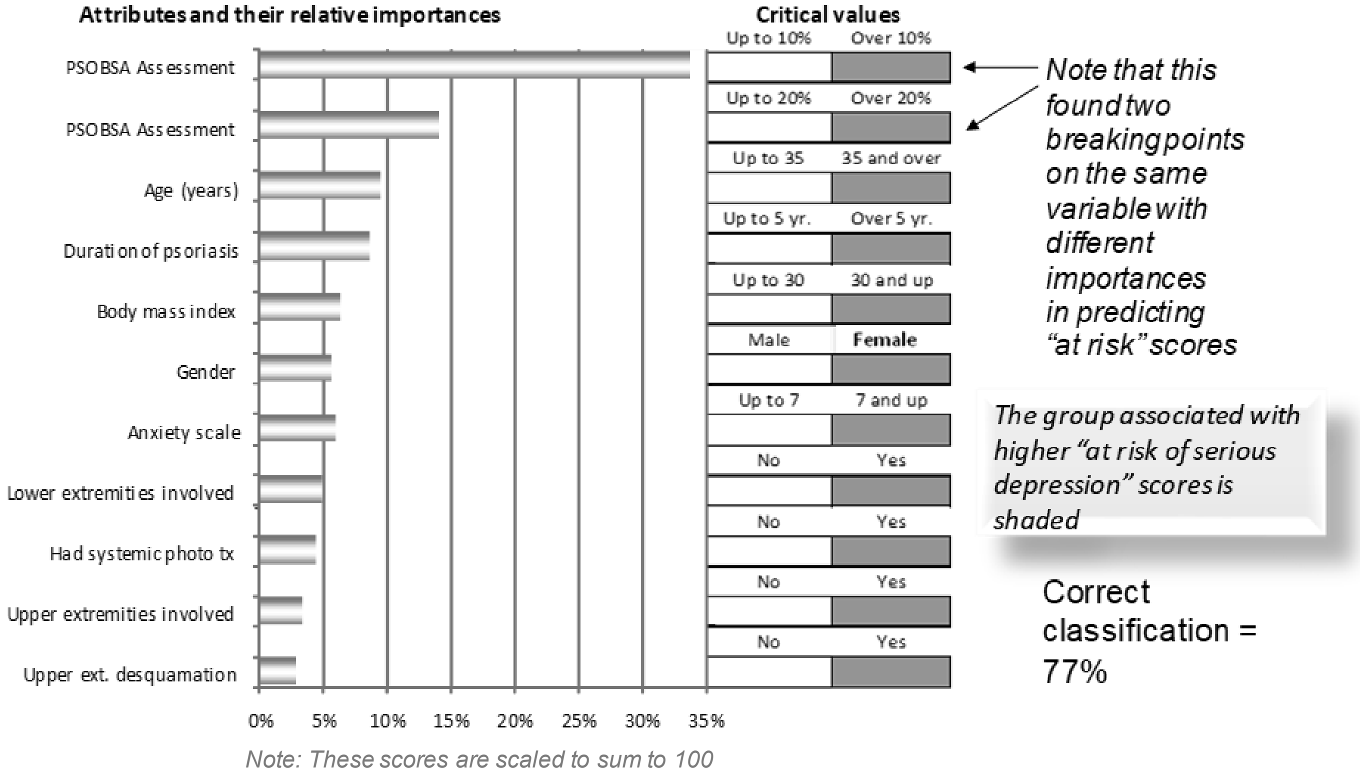
Rather, **boosted decision stumps** first runs a model, a single-level tree (hence the name "stump"). It then learns from that model. The procedure marks which cases are predicted correctly, and which are not. The correct cases are marked as the easy and the incorrect cases as hard. The procedure then puts more weight or emphasis on the hard cases and tries to fit a model that captures them better. It will redo this as many times as you request.

Figure 3 shows an output from a run of boosted decision stumps. This look into variables' importance followed a classification tree model showing the linkages between the nature of psoriasis and depression. The model used measurements of the extent and location of the affected skin areas for about 6,900 patients. These patients also took an internationally-normed test designed to measure serious depression.

The tree model led to a simple set of if-then rules, with each rule corresponding to a different probability of severe depression. It was easy to use, since it was based on measurements that doctors would take in any event as a part of treatment. It could even be scored using a pencil and paper.

When a question arose about the importance of the variables in the model, we suspected that more than one value for some of the variables could be significant thresholds. Therefore, we turned to boosting, which can reveal this type of pattern. The specific method is called **AdaBoost.M1**, which is made to use with classification trees.

We asked the method to run boosting 40 times. As you can see in **Figure 3**, our suspicions were borne out, and two values of the same predictor



Note that this found two breaking points on the same variable with different importances in predicting "at risk" scores

The group associated with higher "at risk of serious depression" scores is shaded

Correct classification = 77%

Note: These scores are scaled to sum to 100

Figure 3: Importances from boosting, relating body measurements to depression

emerged as the most important. This one predictor is **percent of body surface area with psoriasis (PSOBSA)**. One critical threshold is over 10% of body surface area and the other is over 20%. Few other methods match this ability to isolate two values of the same variable as important.

The tree model proved to be valuable because psoriasis patients often conceal how depressed they feel, even from their doctors. The classification tree model gave doctors a simple way to determine which patients might be most at risk. The boosting gave them a few features to watch with extra attention.

We could understand the basic idea of what the method was doing. The specifics, though, remained shrouded in mystery. Other ensembles show us even less. They can produce highly accurate models, and we can apply those models. But otherwise they remain completely opaque.

The ensemble called "decorate"

As mentioned, any basic method for making an estimate can be run repeatedly. We could start with regressions or classification trees or Bayesian networks—or any other approach. There are dozens of ensemble-building methods. Again, some have appalling-sounding names, such as **bagging, dagging** and **MIOptimalBall**.

Decorate is at least a neutral term, and uses an intriguing premise to reach results that are often impressive. By the way, you may sometimes see the term **meta-learners** applied to **ensembles**.

This method uses specially created artificial input data (or **examples**) in making its ensemble estimates. What does that mean? Basically, the method assesses the data using whatever basic analytical method you choose, then creates some number of artificial data cases, or examples, to use alongside the original data in another run of the model. These examples are constructed to maximise predictive performance.

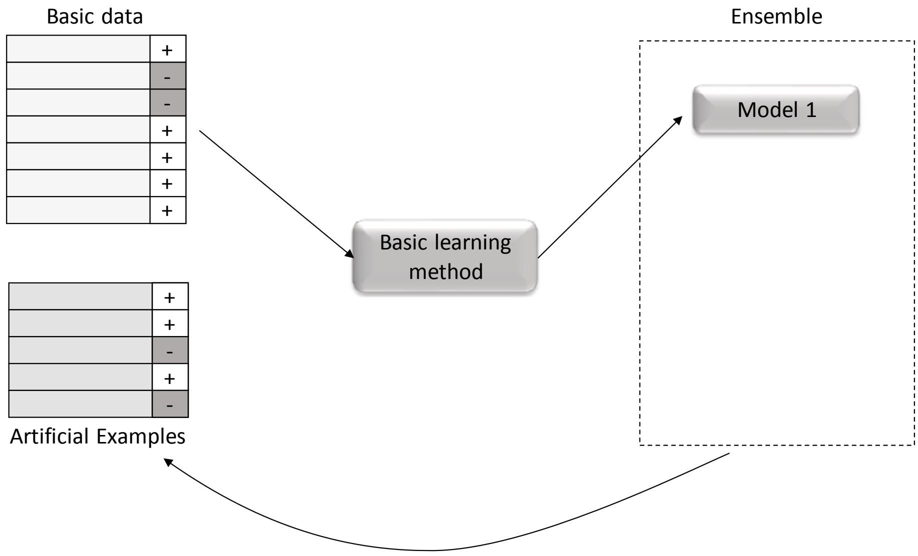


Figure 4: The first step in making a decorate ensemble

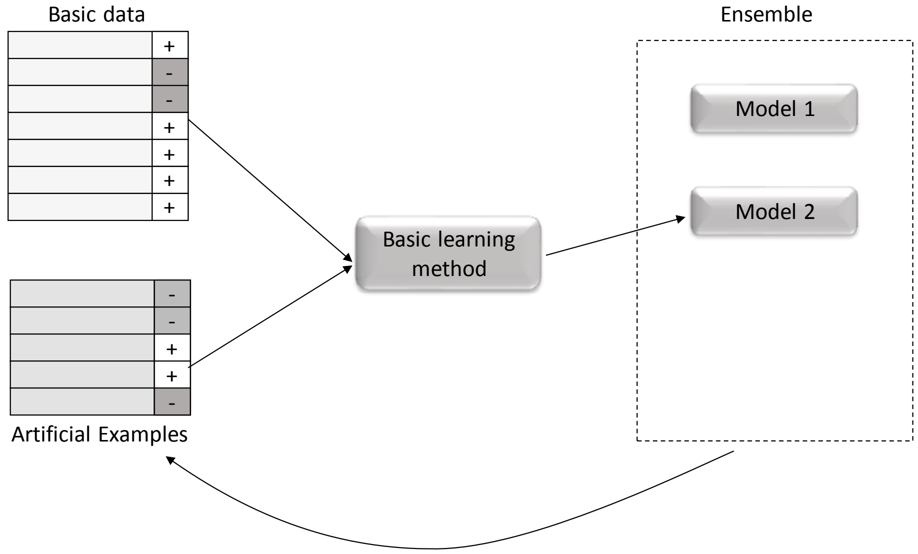


Figure 5: The next step with decorate

This seem like a lot to swallow all at once, so let's take this through a couple of steps. First, the method makes a predictive model, using whatever method you choose.

In this example, we will use classification trees built by a method called **J48**. Although the name may seem unfamiliar, this is a powerful method that can build trees, test them and tear down branches that do not help the overall model—and even relocate branches upward to makes a smaller tree without losing predictive power.

We ran J48, and based on this first run, the program constructed extra data cases or **examples** to add to the original data set. These artificial examples were constructed to boost the predictive accuracy of the model. Specifically, they were made to **disagree** with the basic model. The pooled data therefore is more **diverse** than the original.

Ensembles become useful only if they take a slightly different view of the data each time they run. You may recall that random forests randomly sampled both the predictors that could be in the model and the people in the sample. **AdaBoostMI** reweighted the data, giving more emphasis to the cases that were not classified correctly.

The first step **decorate** takes in changing the pattern in the original data, building artificial data, is represented in **Figure 4**. The artificial cases have a different pattern of plus and minus values than the corresponding cases in the actual data. (Although the box to right is labelled "ensemble," we still have only one model until we get to the next step.)

Once this artificial data set has been made, it and original data are run together. The program now has two models voting. Two models make for the smallest possible ensemble. Most ensembles have many more models.

Based on this run, the program again alters the artificial data. This time, the artificial data is being varied to disagree with what we would expect based on both earlier models. **Figure 5** shows this step.

The program continued. It was set to build 15 ensembles, or to stop earlier if changes in predictive accuracy from one run to another reached zero. You often get more accuracy by running more ensembles.

In this example, we tried a data set that we had already analysed with several other methods, to compare their predictive performance. **Decorate** managed an impressive 91.8% correct prediction level without validation. With validation, this fell to 74.5%, still a very strong level. The best we did with a non-ensemble method was with Bayes Nets, reaching a correct prediction level of 84.0% with no validation and 69.8% with validation.

Decorate therefore did slightly better than the best we could find among many approaches. And this method is often an exceptionally strong performer.

Validation is critical with methods like ensembles, where we cannot check the details of the model. We have to trust that the model has not done anything erroneous—in particular, fitting itself to features peculiar to your data set that do not occur in the outside world. You need a strong validated score before applying the model to other data and using its predictions.

Validation basically involves setting aside (**or holding out**) some part of the data you have, building the model on the rest, and then testing the model on the part that was not used (the hold-out sample). Validation can be done in several ways. The most stringent is called **cross-fold validation**. In this, small parts of the data set are put aside repeatedly, and models are built repeatedly on the remainder. Then results are averaged across all the validation runs.

Decorate pros and cons and ensembles overall

Another strength of **decorate** is that it works well with smaller data sets. This is not always the case for a method using artificial intelligence. For instance, neural nets seem to require great swaths of data to perform at their best. They also may need to be trained and trained, and then trained more. They are getting faster at training, but this is still a real issue where you need an answer to a pressing problem.

Decorate has two drawbacks. First is that it is somewhat slower than non-ensemble methods. All that data construction and learning takes time. For instance, a moderate sized model, using 70 variables and 1800 cases, ran in about 0.1 seconds using Bayes Nets for the initial run and each of the nine cross-validation runs. That is 1 second altogether. It took about 15 seconds with the ensemble.

This is not much of a difference with a data set this size. But when you start thinking of hundreds of variables and hundreds of thousands (or millions) of cases, you will notice the extra time.

We need to reiterate the more salient downside of **decorate**. You cannot see what it has done. It does not even return importance for the variables, and you definitely do not have an interpretable model. You can apply the model to other data, but you must trust it. Again, given the startlingly good predictive performance decorate can reach, you may well decide it is the predictive method of choice.

Still, this impenetrability is the basic shortcoming of all ensembles. If you want to see how the variables fit together, you might well want to stick with a method that provides intelligible visual output, such as the classification trees we discussed or Bayesian networks. Bayesian networks form variables into groups, showing linkages, with the variables having the strongest connections closest to each other. They often perform quite well in predicting outcomes and in revealing strengths of effects.

However, ensembles are always worth exploring, as no method invariably works best in all cases, and they might just outperform your favourite analytical tool in some circumstances. They can also work particularly well if you suspect that some special patterns exist in the data that they alone might reveal, as in our example examining the relationship between depression and body measurements in psoriasis patients.

We spent time with tree-based methods in this paper, because limitations of space (and likely the reader's patience) prevented us from discussing ensembles based on other underlying approaches. However, you can build an ensemble based on any analytical method, such as regression, logit models, or even neural networks. They are a versatile family of methods with many applications. They could well prove to work best for your particular needs.

References

1. Brodley, CE, and Utgoff, PE (1995) Multivariate decision trees, Machine Learning, 19, pp 45-77

2. Buntine, W (1992) Learning classification trees, Statistics and Computing, 2, pp 63-73

3. Clark, LA, and Pregibon, D (1993) Tree-based models, in Statistical Models, eds. JM Chambers and TJ Hastie, pp 377-419, Chapman and Hall, New York

4. Muller, W, and Wysotzki, F (1994) Automatic construction of decision trees for classification, Annals of Operations Research, 52, 231-247

5. Melville P, Mooney, RJ, (2003) Constructing diverse classifier ensembles using artificial training examples, Eighteenth International Joint Conference on Artificial Intelligence, 505-510

6. Struhl, S (2017) Artificial Intelligence Marketing and Predicting Consumer Choice, Kogan Page, London

7. Witten I, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.), Morgan Kaufmann, San Francisco

Author

Dr. Steven Struhl is the author of Artificial Intelligence Marketing and Predicting Consumer Choice (Kogan Page, 2017), Practical Text Analytics (Kogan Page, 2015) and Market Segmentation (AMA Press, 1992, rev. 2013). He also has written over 30 articles on multivariate analysis, computer software, and psychology.

Steven is principal and founder of Converge Analytic. He has 30 years' experience in consulting and research, focusing on applying advanced methods to strategic goals, and framing results and explanations so decision makers can use them effectively. His work addresses how buying decisions are made and understanding consumer groups and their motivations. Earlier experience includes serving 15 years as Senior Vice President, Senior Methodologist at Total Research (later Harris Interactive). He also served as Director of Market Research at SPSS, Inc., where he guided development of new statistical software. Before that, he held senior positions at Harris Bank/Bank of Montreal, and in advertising at Draft/FCB.

Steven speaks frequently at conferences and has given numerous seminars on pricing, choice modelling, market segmentation, text analytics and presenting data; has taught graduate courses in statistical methods and data analysis, and is now teaching online certification courses in data analysis and statistics.

He holds an MBA from the University of Chicago Booth School of Business, a doctorate in psychology from the Chicago School of Professional Psychology, and MA and BA degrees from Boston University.

CATALINA®
Data & Analytics
Relevant. Real Time. Results.
Catalina helps you **optimize your marketing** so you **sell more** and **spend less.**
Driving insights with a data set you can't get *anywhere* else.
buyR³ science™
Relevant. Real Time. Results.

FRONTIERS OF MARKETING
DATA SCIENCE JOURNAL

ISSUE No. 1 | 2018
www.i-com.org