USING CROWDSOURCING AND IMAGE PROCESSING TO AUTOMATE DATA COLLECTION

By Tamas Gaspar, Principal Statistician
Siew-Sim Lim, Associate Director, Buy Methods, Nielsen
The world of measurement is changing.

Thanks to recent advances in data collection, transfer, storage and analysis, there’s never been more data available to research organizations. But ‘Big Data’ does not guarantee good data, and robust research methodologies are more important than ever.

Measurement Science is at the heart of what we do. Behind every piece of data at Nielsen, behind every insight, there’s a world of scientific methods and techniques in constant development. And we’re constantly cooperating on ground-breaking initiatives with other scientists and thought-leaders in the industry. All of this work happens under the hood, but it’s not any less important. In fact, it’s absolutely fundamental in ensuring that the data our clients receive from us is of the utmost quality.

These developments are very exciting to us, and we created the Nielsen Journal of Measurement to share them with you.

WELCOME TO THE NIELSEN JOURNAL OF MEASUREMENT

SAUL ROSENBERG

The Nielsen Journal of Measurement will explore the following topic areas in 2016:

**BIG DATA** - Articles in this topic area will explore ways in which Big Data may be used to improve research methods and further our understanding of consumer behavior.

**SURVEYS** - Surveys are everywhere these days, but unfortunately science is often an afterthought. Articles in this area highlight how survey research continues to evolve to answer today’s demands.

**NEUROSCIENCE** - We now have reliable tools to monitor a consumer’s neurological and emotional response to a marketing stimulus. Articles in this area keep you abreast of new developments in this rapidly evolving field.

**ANALYTICS** - Analytics are part of every business decision today, and data science is a rich field of exploration and development. Articles in this area showcase new data analysis techniques for measurement.

**PANELS** - Panels are the backbone of syndicated measurement solutions around the world today. Articles in this area pertain to all aspects of panel design, management and performance monitoring.

**TECHNOLOGY** - New technology is created every day, and some of it is so groundbreaking that it can fundamentally transform our behavior. Articles in this area explore the measurement implications of those new technologies.
To measure the consumer packaged goods retail trade, marketing research companies such as Nielsen typically collect data directly from retailers who provide electronic point-of-sale (POS) information from their check-out scanning systems. This is by far the most accurate data available, but its collection is dependent upon retailer cooperation. If some of the retailers in the sample design do not cooperate, there can be some level of bias in the reported data. We can eliminate that retailer cooperation dependency by collecting data directly from individuals in population-projectable consumer panels, such as Nielsen’s consumer panel services (CPS), but the size of panels in certain regions of the world makes it difficult sometimes to report data at the granularity clients need to track performance. Increasing sample size is a solution, but it’s not always possible due to the costs of managed panels and the difficulty of recruiting reliable panelists.

With the growing worldwide adoption of new technologies like mobile smartphones, crowdsourcing, and virtual payments, new opportunities now exist to collect purchase information directly from consumers, in large numbers, and do so in a way that is both economical and less burdensome.

*To operate its flagship consumer panel services, Nielsen provides each panelist with a portable scanning device and requires that they scan the bar code of each and every product they purchase.*
KEY MARKET CATALYSTS FOR AUTOMATION

According to Gartner, the smartphone market today has reached 90% penetration in the mature markets of North America, Western Europe, Japan and parts of Asia/Pacific⁴. More than 1.4 billion smartphone units were sold worldwide in 2015⁵. It’s become second nature for people to use apps on their mobile phones, and to use their mobile phones to share content over the internet. With this rapid adoption of smartphones by consumers comes opportunity for data collection. The quality of built-in cameras in the new generation of smartphones has improved to the point where images taken with those phones have a good enough resolution to serve as input data for automated processing. Asking people to take pictures of their store receipt is the basis for the project we’re describing in this paper.

People are not as reluctant to participate in cooperative projects as they might have been in the past. In fact, the rise of modern technologies and social media is helping shine a bright new light on the world of crowdsourcing⁶. Without the technological barriers and social stigma, the notion of recruiting people to take pictures of their store receipts with their phones is not far-fetched anymore. With proper engagement and motivation, we can use the crowd very efficiently. Today’s online volunteers are very comfortable with digital rewards. No need to deal with physical goods or currency anymore: We can reward participants with “virtual coins,” and use the internet to transfer and manage those rewards.

To test these assumptions, we ran a proof-of-concept study in the United Kingdom. The U.K. was an ideal test market for two reasons: First, the overall smartphone penetration in the U.K. at the start of 2016 was around 68% (close to 91% for people under 35 years old, and 60% for people over 35)⁷. High enough for reasonable representation, and with improvements for better representation of the whole population expected in the near future. Second, we estimate that the combined market share of the two big non-cooperating discounters in the U.K. (Aldi and Lidl) is over 10%, so there is real market demand to provide more precise insights about these chains.

TRYING OUT THE NEW PROCESS IN THE U.K.

We kicked off our proof-of-concept project in January 2016, and by the end of August, 8,000 users had signed up and uploaded at least one receipt from their camera phone. So far, we have collected more than 400,000 images. In the first stage of the project, we focused on process and the best ways to engage with participants. We started with a small-scale initiative with approximately 800 users contributing pictures of their receipts back in January, and planned for a gradual increase to 4,000 active users by the end of the year.

Through random web-based invitations, we ask users to install our app on their smartphone. During the sign-up process, we collect some basic demographic information: their age, household size and postal code, as well as their email address. This information is used later in the process to measure and control for bias. The app is really simple: Users take a photo of any purchase receipt they get when they buy any product in a store. The photo is taken from within the application (that is, not with the phone’s standalone camera app), so that the app may immediately assess whether the quality of the image is up to our standards, and alert the user to retake the shot if needed. A typical receipt includes the description of all purchased items, the date and time of the purchase, the street address of the shop, as well as the amount paid. Each image is sent to the cloud for us to access, download and process the information further.

The user is rewarded with virtual coins for uploading even a single receipt. Additional coins can be earned if the app is used daily, or if the user invites others to become participants. Our objective isn’t to recruit a managed panel with this project, hence the way we recruit participants (using referrals, for instance) and the structure of the reward system itself are not subjected to the same scrutiny. We do want the rewards to be motivational, but of small value, so that they don’t unduly influence the shopping behavior of our participants. The virtual coins that people earn while participating in the study can be redeemed to enter a sweepstake, or saved up and eventually cashed out. There’s no minimum that needs to be achieved before earning the right to take part in a sweepstake, so a user may have just signed up, uploaded one receipt, and already earned the right to bid on a reward. The more coins users bid on a reward, the better their odds of winning.

---

⁴http://www.gartner.com/newsroom/id/3339019
⁵http://www.gartner.com/newsroom/id/3215217
⁶Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people (like an online community) rather than from employees or suppliers
⁷Pew Research Center - http://pewrsr.ch/VRXj4q
While we instruct users to take pictures of all their receipts, we cannot be certain that they’re complying. Perhaps they’re only sending us a few when convenient, or occasionally submitting receipts from other shoppers (e.g., other household members or even neighbors) in order to earn more rewards. To detect these situations, we have developed statistical algorithms to determine probabilities that users might under- or over-deliver. To maximize the usefulness of the receipts we receive, we are looking at them as being representative of the universe of all shopping trips in the country to a specific chain. This is less subject to bias, but it also means that we need to establish a good independent estimate for the total number of shopping trips to each chain during the reported period.

INITIAL PARTICIPATION BUILD-UP

This is a proof-of-concept, and it was important for us to focus on the mechanics and logistics of the project rather than make sure that the participants were absolutely representative of the population. The initial phase allowed us to build quality checks to detect users whose activity was not compliant with our request (e.g., sending an image of a product instead of a receipt, or sending an obsolete receipt, or a receipt already uploaded by someone else) and implement ways to warn those end users. Users who repeatedly broke the rules were suspended, and not allowed to redeem any reward. We also learned that users stayed active on average only one month, so we started to experiment with ways that we could extend the duration of their commitment. Early data results have confirmed our initial hypotheses about some bias areas: We have low-representation among the oldest population group (they are the least technology-savvy of all population groups), and the average basket size is slightly less than what we originally estimated. We expect to be able to counter-balance these biases by using our managed household panel data for calibration.

For efficiency, we process all receipts via a sophisticated optical character reading (OCR) solution. OCR comes with great automation benefits, but it’s also an enormous technological challenge in our particular situation. One complication is dealing with a large volume of images with noisy background, and developing an algorithm to remove that noise is not a trivial endeavor. Another difficulty is to properly transform each receipt image into a set of database entries. The algorithm needs to be capable of finding the relevant information on each receipt; interpret each data point accurately regardless of its position on the receipt; and determine the correct meaning of the product descriptions found on that receipt.

IMAGE CAPTURE AND ITS CHALLENGES

It’s critical that the quality of each image satisfies a minimum-level requirement, and machines have yet to catch up with the human eye’s capability of deciphering text that might be blurry, printed on crimpled paper (e.g., warped, wrinkled, partly folded), in uneven lighting, or with faded or missing characters. Additionally, receipts might show up placed at an angle on the picture, or with objects in the background, or they could even be of an object that is not a receipt to begin with. Think about online CAPTCHA systems: To prove you’re not a bot, those systems present you with strings of blurry, distorted characters that are typically easy for a human being to make sense of, but nearly impossible to decipher for a bot relying on OCR algorithms. We want to prevent a situation where the store receipts we receive are as unreadable as those CAPTCHA string challenges.

For OCR to be successful with pictures of paper receipts, uploaded images need to be de-skewed, with as little background noise as possible, and the print needs to be clear. Image quality at the source is an all-important determinant for the success of OCR processes. We have a user guide to assist users in placing the camera in an optimal position, and to instruct them to take pictures of their receipts (even long, multi-part receipts) with as little background noise as possible. As long as users follow these instructions, there’s no need for them to perform more complex operations—such as cropping the image or any other special editing.

Taking the OCR out of the phone and putting it in the cloud was an important design decision. The Nielsen app can ascertain picture quality directly on the phone and alert the user to retake the picture if needed, but the more complicated image processing steps are performed out of the phone. There are many different types of smartphones in circulation, and only high-end mobile phones have enough processing power to perform sophisticated image processing and OCR with any reasonable degree of success. We decided to develop

4 For an illustration of how big data can effectively be calibrated using panel data, see “The value of panels in modeling big data” by Paul Donato, Nielsen Journal of Measurement, Vol 1, Issue 1 (July 2016)
our own app so that our success rate wouldn’t be a function of the type of smartphone device used by participants. Taking the OCR out of the phone also allowed us to try new detection algorithms with much more flexibility than if we had to upgrade the app on everyone’s phone every time we wanted to change something.

PREPARING THE IMAGES FOR OCR

Once an image reaches the Nielsen server, the machine work begins. The first checkpoint is that of image quality. The image has already passed a first test on the phone, but we’re able to apply more advanced algorithms on the server to detect blurriness and determine if a receipt is good enough for OCR. If it passes that test, our next step is to identify the store chain via a combination of logo and character recognition techniques. We built a dictionary of the most frequent logos and their variants so that the tool can match the image to the ones in the dictionary. A number of innovative patent-pending techniques are at play to extract relevant regions of interest on the image, detect a logo and identify the source of the receipt.

If the image cannot pass those tests, it’s siphoned off to a manual intervention stage where someone takes the necessary steps—like manually cropping the region of interest before sending the image to the de-skewing algorithm, or visually identifying the chain before sending the receipt back for further processing. When we started the project in the U.K., half of the receipts at that stage required some level of manual intervention, but the experience built over the course of the last few months has allowed us to bring that figure down to 25%, with plenty of room for further improvements.

NIELSEN’S UNIQUE OCR SOLUTION

Nielsen’s patent-pending innovation consists of wrapping the OCR engine with a manual correction function that uses machine learning to teach the function itself to autocorrect going forward. Effectively, what this means is that manual corrections on a small percentage of images end up having a much larger impact, as learnings are gained by the machine and applied to future batches. Early tests indicate a 90-95% recognition rate based on this machine learning mechanism.

A recent run showed an impact rate of 22 times the size of the manually corrected batch: that is, for every receipt that underwent manual correction, 22 similar receipts benefited from the learning and were corrected automatically.

Downstream, the next step is to make sense of the text that has come through. We’re using a specialized algorithm to classify the recognized text into relevant data types for the business (e.g., product name, quantity, price, etc.). The information needs to correspond to products that have already been inventoried by the system—or if the product is not recognized, invite a manual intervention to add the product to the database.

A number of things could still go wrong in that late stage. For instance, it may be difficult to determine whether a new product is indeed a new product, or an existing product with an alternate spelling. We may also find a discrepancy between the sum of values of all products on the receipt, and the total sum at the bottom of the receipt. As we are still in the proof-of-concept stage, the data dictionaries are still a work in progress. We’ve found that many products appear only on one receipt, and the large number of such products poses a special challenge to the team responsible for cross-coding these descriptions. The maintenance of the product and store dictionaries is one of the largest cost components in the whole production process. But it’s well worth the investment: Our team recently discovered that many receipts from one of the retailers in the study weren’t getting processed properly because the chain had just switched to a new logo. Once that correction was made in the visual library, the recognition rate for that retailer jumped by 35%.

PUTTING IT ALL TOGETHER

There are many different pieces to this process: collecting the images, performing image quality checks, identifying the retail chain, extracting the relevant regions-of-interest, performing the OCR itself, and finally classifying the related text—calling upon manual intervention along the way, as needed. In order to keep track of everything and follow every image along its life cycle, we built an end-to-end web facility that all team members can use to measure progress.
The illustration above provides a glimpse of the steps involved and their current success rates for one of the chains in the study. Out of every 100 images captured and uploaded by participants, 85 currently pass our initial quality checks. The 15 images that don’t pass this early check are either incomplete, unreadable, duplicates or are not receipts to begin with. Of the 85 that pass through, 17 correspond to long receipts that we don’t currently handle automatically, although that’s an active area of development. The 68 images that correspond to single receipts are then prepared for OCR processing, and 75% of those (i.e., 51 images) make it through to the actual OCR stage. Of those, 48 are successfully processed and only three need to be checked.

These early results are very encouraging, but there's plenty of room for improvement. Our success rate keeps climbing from month to month as discoveries are made regarding the nature of the receipt images flooding in, and as our software stack becomes more sophisticated to address new scenarios.

CONCLUSION

The proof-of-concept study currently underway in the U.K. is an outstanding opportunity for Nielsen to test the viability of a new paradigm in data collection. It doesn’t replace point-of-sale scanner data, of course, but it provides an effective method to fill in the gaps when the point-of-sale data in a particular market doesn’t quite cover the universe of retail outlets. For our CPS managed panels, the study opens the door to potentially transition collection from Nielsen proprietary handheld scanners to mobile apps that lessen the requirement of scanning each item purchased to simply capturing an image of the overall receipt. Receipts come in many forms, but our algorithms are getting smarter all the time at handling special cases. While our software development effort is far from finished, we already have enough confidence in the quality of the automated outcome to start working on potential rollout plans and expanding this work to other countries.