Submitted to *Management Science*
manuscript MS-13-00168.R3

Cultivating Disaster Donors Using Data Analytics

Ilya O. Ryzhov
Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, iryzhov@rhsmith.umd.edu

Bin Han
Applied Mathematics, Statistics, and Scientific Computation, University of Maryland, College Park, MD 20742, danielh@math.umd.edu

Jelena Bradić
Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, jbradic@ucsd.edu

Non-profit organizations use direct-mail marketing to cultivate one-time donors and convert them into recurring contributors. Cultivated donors generate much more revenue than new donors, but also lapse with time, making it important to steadily draw in new cultivations. The direct-mail budget is limited, but better-designed mailings can improve success rates without increasing costs. We propose an empirical model to analyze the effectiveness of several design approaches used in practice, based on a massive dataset covering 8.6 million direct-mail communications with donors to the American Red Cross during 2009-2011. We find evidence that mailed appeals are more effective when they emphasize disaster preparedness and training efforts over post-disaster cleanup. Including small cards that affirm donors' identity as Red Cross supporters is an effective strategy, while including gift items such as address labels is not. Finally, very recent acquisitions are more likely to respond to appeals that ask them to contribute an amount similar to their most recent donation, but this approach has an adverse effect on donors with a longer history. We show via simulation that a simple design strategy based on these insights has potential to improve success rates from 5.4% to 8.1%.

Key words: business analytics; non-profit operations; donor cultivation

History: Received January 24, 2013; accepted December 18, 2014, by Serguei Netessine, operations management.

1. Introduction

When a major disaster strikes (e.g. Hurricane Katrina or the Haiti earthquake), non-profit organizations experience sharp spikes in one-time donations. These donations are coordinated for immediate disaster relief, as well as a wide variety of "development" programs, such as community disaster preparation, emergency response training, and sustainability efforts. However, fewer than 30% of one-time donors return to give a second time. The unpredictability of donor response limits managers' ability to plan long-term operations for programs that require steady funding (Tomasini and Van Wassenhove 2009). In order to secure a consistent, reliable cash flow, non-profits devote significant efforts toward cultivating one-time "disaster donors" into long-term donors.

The importance of donor cultivation is widely acknowledged throughout the non-profit industry. Fundraising guides (Burnett 2002, Klein 2007) emphasize the distinction between recurring (or "warm-list") and completely new donors. Managers at the American Red Cross have found that warm-list donors are 4-5 times more likely to respond to direct mail solicitations; that is, the response rate of a cultivation program may be 5%, whereas a program designed to acquire completely new donors may achieve a 1% success rate. A survey by Sargeant et al. (2006) of 150 UK-based charities reports similar numbers. In fact, the short-term revenue (gifts minus costs) of acquisition programs is often negative, and most donors only begin to generate revenue after cultivation (Sargeant and Kähler 1999).

Cultivation is largely accomplished through direct mail: Sargeant and Jay (2004) observes that direct-mail donors "exhibit higher levels of loyalty" and generate greater lifetime value than those targeted by other forms of direct marketing, while Sargeant et al. (2007) notes that donors that are acquired online are often subsequently cultivated through direct mail. For the Red Cross, direct mail accounts for about 2/3 of the total direct marketing budget, and 85% of gifts collected through mail are made by recurring donors. However, fewer than 50% of these are retained from one year to the next, so it is important to ensure that new donors are always being converted. Simply sending more mail may not be an effective way to achieve this goal (Meer and Rosen 2011), and in any case may not be feasible under a fixed budget.

In this paper, we study the problem of improving conversion and retention rates through the design attributes of the mailings themselves. "Design" can refer to the content of the letter, the presence or absence of a free gift (such as a set of address labels), various methods for setting suggested donation options, and other marketing strategies. Both practical fundraising guides (Warwick 2008) and academic studies (Bekkers and Wiepking 2010) point to connections between the design and the outcome of a fundraiser. There is a large literature on social, psychological, and economic mechanisms driving charitable donations, which could potentially be leveraged by various design strategies. For example, the warm-glow model (Andreoni 1990) may be used to lend support to letters that elicit a positive emotional response in the donor, while the gift-exchange model (Falk 2007) may suggest a more direct incentive (such as a gift item). There is also a substantial experimental literature on charitable giving, some of it involving specific fundraising practices such as matching grants; see Sections 2-3 for examples.

However, while this work provides numerous insights into, e.g., donor behaviour, it is less clear how these insights can be operationalized. Empirical work in this area often considers data that are voluntarily provided by donors, or collected from lab experiments, where individual behaviour can be closely observed. However, in everyday practice, a non-profit manager may have records of millions of past communications with donors, but does not have access to detailed individualized information. The operational decision of how to design and target a new fundraiser must be based on the information that is visible to the organization. Further complicating the issue, the literature frequently does not distinguish between donor acquisition and donor cultivation. Some studies (Diamond and Gooding-Williams 2002, Karlan et al. 2011) observe clear differences in how warm-list donors respond to appeals compared to new donors. Still, it is not clear how these differences should affect managerial decisions; furthermore, as Meer and Rosen (2011) observes, experimental insights may be difficult to corroborate empirically using observational data.

The present paper seeks to close this gap. Our main contribution is an empirical analysis that identifies design attributes of direct-mail fundraisers that exert a significant impact on donor cultivation and retention. The context for our analysis is provided by a dataset jointly compiled by the Red Cross and Russ Reid Company during 2009-2011, for a cultivation program known as STAART (Strategy Through Applied Analytics, Research, and Testing). The dataset covers $49 million in donations to STAART from over 300,000 donors, with detailed campaign information available for over 8,000,000 individual recorded communications with over 1,000,000 individuals. Specifically, we have records of the characteristics of the outreach strategy used, such as the design or formulation of the mailed appeal; limited characteristics of individual donors, such as their previous donation amounts; and some characteristics of disasters such as their magnitude and location. We also consider *interactions* between design attributes of campaigns, thus accounting for potential heterogeneity of design effects by donor class.

We focus on four important design strategies used by the Red Cross; our insights into their effectiveness also shed light on the distinct nature of cultivation campaigns. First, Red Cross mailings can include small "supporter cards" affirming the donor's identity as a Red Cross supporter. These cards are included in only 5% of mailed appeals, but our results suggest that they have a significant positive effect on the donor's likelihood of returning. Second, we study different ways of composing the "story," or the written appeal included in a mailing. We obtain the surprising result that the most effective stories focus on disaster preparedness rather than relief efforts, suggesting that informative rather than emotional content may be more useful in cultivation campaigns. Third, we find that free gift items (such as address labels) included in Red Cross mailings appear to make no significant contribution to donor cultivation. Finally, we consider the practice of dynamically generating the suggested donation amounts based on the donor's previous behaviour (e.g. 75%, 100%, or 150% of the donor's most recent donation). We find evidence that this technique improves conversion of recent new donors, but has an adverse effect when targeted at recurring or lapsed supporters. Finally, we use simulations to show that a simple design strategy based on all of these insights has the potential to increase response rates from 5.4% to 8.1%, suggesting that our analysis translates into useful operational recommendations.

To summarize, we contribute to the literature on non-profit analytics by identifying designs that exert a significant impact on the outcome of a fundraising campaign, as well as key interactions between these designs and various donor segments. To our knowledge, this is the largest study to date on the interactions between *donors*, *disasters*, and *designs*. Our results (e.g., for preparedness stories and gift items) suggest ways in which cultivation campaigns (such as STAART) should be considered differently from other types of fundraisers. These insights lead to clear, simple policy recommendations; we conduct simulations that suggest that these recommendations have significant potential to improve fundraising efficiency.

2. Hypotheses development

The main objective of this study is to identify design elements of direct-mail fundraising campaigns that exert a significant effect on donor cultivation. Below, we present four research questions, each associated with a particular design element in the data. These design elements are supporter cards, written appeals, free gift items, and dynamic donation amounts.

1. *Supporter cards.* Approximately 5% of Red Cross mailings contain a "supporter card," a small card bearing the donor's name, identifying the donor as a Red Cross supporter, and stating the organization's mission. Donors are encouraged to carry their cards with them. The cards cannot be redeemed for money or gift items; their value lies solely in confirming the donor's support of the organization. The question of interest is whether these cards contribute significantly to donor retention.

We hypothesize that cards will positively impact cultivation. The motivation for this hypothesis relates to the concepts of self-identity and social identity in social psychology. Essentially, this literature argues that donors will be more likely to give if it is an important part of their conception of who they are (self-identity), or if it is closely associated with membership in a certain group (social identity). These concepts have been previously studied in connection with charitable behaviour. For example, a pioneering study by Charng et al. (1988) found empirical evidence suggesting that self-identity plays a significant role in the intention to give blood. It was also found that self-identity was more highly correlated with *repeated*, rather than one-time, charitable actions. This is particularly noteworthy for us, since we focus on donor cultivation rather than on attracting new first-time donors, and our dataset consists entirely of donors who have already given once.

More recent work has considered the role of self-identity (measured, e.g., using questionnaires or surveys) in bone marrow and organ donation (Simmons et al. 1993, Hyde and White 2009), recycling (Terry et al. 1999), and ethical consumption (Shaw et al. 2000). The empirical evidence suggests that self- and social identity are strongly correlated with charitable intentions. These findings support the hypothesis that the Red Cross supporter card, whose sole value derives from its ability to reinforce the donor's identity (both individual and group) as a Red Cross supporter, is helpful in eliciting recurring donations. Note that this is a managerial question, not a behavioural one; for example, the question of whether supporter cards increase the proportion of people who identify as Red Cross supporters, or whether they help draw a response from those who already identify as such, is a separate problem outside the scope of our study (and our data). From the manager's point of view, the key insight is whether cards exert a positive impact on donor retention (thus implying that they are highly under-utilized).

2. *Formulation of appeals.* Red Cross mailings typically formulate the appeal for donations (the "story") in one of three ways. A *specific* disaster story describes relief efforts after a particular, named disaster; a *generic* story describes similar efforts without reference to a specific event; and a *preparedness* story describes various community services that will be offered to victims of future disasters (e.g., "The local chapters of the Red Cross must be prepared to provide food, shelter and counseling for disaster victims...support firefighters and other emergency personnel...[and] they must provide CPR, first aid and other emergency preparedness training..."). The organization wishes to know which of these three approaches is the most effective.

A recent empirical study of charitable giving by Karlan et al. (2011) found that an urgent message that was not accompanied by a clear reason actually was harmful to the outcome, suggesting that generic stories may be the least effective of the three. With regard to specific vs. preparedness stories, there exists a volume of empirical evidence to support the hypothesis that specific stories are more effective, due to their more visceral and emotional content. For instance, Bennett and Kottasz (2000) found that "highly emotive advertising imagery" played a major role in convincing people to contribute. Similarly, Bennett (2009) found that emotional, rather than informative, content was more likely to elicit online donations, whereas Small and Loewenstein (2003) found that donors were more likely to contribute when they could identify their donation with an individual victim, even if no personalized information about the victim was provided. From this viewpoint, a preparedness story is especially non-identifiable, since at most it can describe potential future victims, not existing ones.

Counter-arguments are also available. In a survey by Breeze and Dean (2012), prospective donors expressed the opinion that, even if an appeal identifies a specific recipient of charity, the donor may not be able to relate to that recipient. Visceral disaster imagery may also be perceived as exploitative or distasteful (Calain 2013). Nonetheless, even the donors surveyed by Breeze and Dean (2012) hold to the conventional wisdom that emotional appeals work.

One important nuance, which is particularly relevant for our study, is that studies such as Small et al. (2007) and Bennett (2009) primarily consider *impulsive*, spur-of-the-moment decisions made by first-time donors. By contrast, we consider the problem of donor *cultivation*; the donors in our dataset have already made a first-time decision to give, and the task is to convince them to become regular supporters. In fact, Bennett (2009) observes limited evidence that "knowledgeable and committed givers" are more likely to respond to informative, rather than emotional, content. The distinction between the various types of stories is thus especially important for programs that focus on donor retention and cultivation.

3. *Gift items.* Occasionally, Red Cross mailings include free gift items, such as address labels or emergency lights. Since these items increase costs, we would like to know if they actually improve the likelihood of repeated donations.

The literature on gift incentives for charitable donations is quite mixed. On one hand, Landry et al. (2006) found that the offer of a chance to win a lottery prize positively influenced giving, while Falk (2007) found that a set of free postcards positively influenced responses in a direct-mail campaign. However, Gneezy and Rustichini (2000) notes that the impact may depend on the value of the gift ("small" or low-value gifts may have an adverse impact, even if high-value gifts are effective). Frey and Oberholzer-Gee (1997) provides theoretical and empirical arguments that compensation may adversely affect "intrinsically motivated" participants who are driven by altruism or a sense of civic duty. Landry et al. (2010) finds an important nuance: while new donors do seem influenced by gifts, this strategy may be ineffective for "warm-list" donors who have given to the organization before. For this reason, we hypothesize that gift items will *not* be effective, since every donor in the STAART dataset is a warm-list donor.

4. *Dynamic donation amounts.* A major question in the design of Red Cross appeals is how much money to ask for. In 45% of mailings, these amounts are determined dynamically based on the donor's past behaviour. For instance, the donor may be given options to donate amounts equal to 75%, 100%, or 150% of that individual's most recent donation. The appeal gives the amounts, not the percentage values. The question is whether this strategy is more or less effective than asking for preset, "standard" amounts.

To our knowledge, this specific question has not been previously studied in the literature. Doob and McLaughlin (1989) experimented with fixed amounts of different magnitudes, and recommended simply asking for a large (though still realistic) fixed amount. By contrast, Desmet and Feinberg (2003) considered several scales for generating dynamic amounts, but did not compare against a benchmark, non-dynamic strategy. For our purposes, the most relevant result from this study was that past donation amounts are a significant predictor of future donations, suggesting that dynamic amounts may be a reasonable strategy. A recent field experiment by Shang and Croson (2009) finds that increased contributions in one year do not crowd out future contributions, also suggesting that donors may be willing to replicate their past behaviour if given the chance. However, it is not clear whether this effect may be heterogeneous for different donor segments (e.g. donors whose first contribution occurred recently, vs. those who have already given multiple times). We also consider such segment-specific effects in our analysis.

3. Other related work

The subject of charitable donations has been studied in economics, marketing, and public policy. There has been relatively little work focusing on donor cultivation and retention, particularly from the operational perspective; we believe our work to be the first large-scale study of this type. Below, we survey the viewpoints of other communities and contrast them with our own.

Both theoretical and empirical studies of non-profit donations have often focused on the impact of donor income level, donor demographics, and policy decisions such as tax credits, on trends in charitable giving. Empirical work on this topic tends to rely on publicly available surveys on family expenditures (Jones and Posnett 1991, Kitchen 1992), which offer detailed income and demographic data on relatively small samples of households. One example of this demographics-oriented approach is a recent study by Brown et al. (2012), based on the U.S. Panel Study of Income Dynamics, which surveys 5,000 families in the United States (PSID 2012). See also List (2011) for a thorough demographic analysis of the market for charitable donations. Other recent work has presented evidence correlating donation amounts with other factors such as media coverage (Brown and Minty 2008) and trends in the stock market (List and Peysakhovich 2011).

The operations perspective has mostly considered revenue management and efficient resource allocation (de Véricourt and Lobo 2009, Privett and Erhun 2011, Lien et al. 2014). A recent work by Leszczyc and Rothkopf (2010) studied how charitable motives can maximize the revenue generated in an auction. The literature also contains a number of theoretical models designed from the donor's point of view, e.g. seeking to optimally allocate resources to maximize a utility function (Yen 2002, Landry et al. 2006). There is also a great deal of interest in behavioural drivers of donations. For example, Arnett et al. (2003) studies how the prestige of a university affects alumni donations. Fennis et al. (2009)

conducts lab experiments to study how "foot-in-the-door" behaviour (e.g. asking potential donors to fill out a survey before asking them to donate) affects the likelihood of donor response, while Shang et al. (2008) studies how donor motivation is affected by information about other donors. Van Diepen et al. (2009) considers the propensity of repeated direct mailings to irritate donors and negatively impact retention, whereas experiments by Karlan and List (2007) and Karlan et al. (2011) study how donations are impacted by the promise of matching grants. Donor motivation is another important topic of research (Andreoni 2006, Schokkaert 2006), but is outside the scope of our study (and our data). We use the data to infer variation between donors, whether it stems from behavioural, demographic, or economic factors.

The literature has considered a number of theoretical econometric models for predicting donation amounts. One widely-used class of such models is known as the Tobit model (Lankford and Wyckoff 1991), applied e.g. by Auten and Joulfaian (1996) to investigate the effect of income and estate taxes on donations. This approach (see also the extension in Brown et al. 2012) is motivated by the particular structure of donations, which are always non-negative, and have a high incidence of zero values, because many surveyed households do not give to charity at all. In our setting, however, most individual donations to the Red Cross are fairly small, and the organization places high value on the *incidence* of donation (that is, the ability to reliably elicit a response), as opposed to the monetary amount. Additionally, while the organization can distinguish between individual donors, it does *not* have access to personally identifiable information (PII) about them (e.g. income or demographics). We rely on the data to establish the drivers of donor cultivation, treating the PII as an unobservable random effect.

The Red Cross has been the subject of extensive previous study. For example, Lacetera et al. (2012) studies incentives in the context of blood donations, while Lafferty et al. (2004) considers consumer attitudes toward brands that partner with the organization. Behavioural studies such as Ariely et al. (2009) have used Red Cross donations to provide a context for studying donor motivations. Logistical issues faced by the organization have been studied e.g. by Pedraza-Martinez and Van Wassenhove (2013). To our knowledge, however, our study is the first to focus on the effective design of direct-mail fundraisers, particularly in regard to donor cultivation. Our work is closer to Bult et al. (1997), which also studies the incidence of donation for a direct-mail fundraiser (by a different organization), with a dataset covering 48,000 communications with 3,000 households, and a small number of basic design attributes such as the presence of a brochure or payment slip in the mailing. By contrast, we study a massive dataset with over 8 million communications and a rich set of donor and campaign attributes (up to 300 in the largest model we consider), which enables us to draw detailed conclusions about the efficacy of the design elements described in Section 2.

4. Description of STAART data

New donors enter the STAART database by contributing to a specific disaster relief campaign (e.g. after a major disaster), or by responding to an acquisition campaign. These new *acquisitions* subsequently receive mailed appeals encouraging them to continue their support. If a donor responds to such an appeal, he or she is said to be "converted," and is considered a current supporter of the program. Note that, in order to be converted, a donor must give at least two donations: the first donation identifying the donor as an acquisition, and a second donation in response to a conversion attempt. Once converted, a donor is regularly sent several types of mailings, broadly classified as *renewal* (direct appeals for a contribution), *cultivation* (newsletter-like mailings, primarily intended to build a relationship with the donor) and *follow-up* (other intermittent mailings). If the donor does not respond to any of these appeals for a period of 18 months, he or she is reclassified as *lapsed*. Some campaigns have a *generic* type, meaning that they are catch-all fundraisers targeted at all donors.

The Red Cross dataset consists of several large lists that separately catalogue communications, donations, donors, and disasters. We performed data processing to cross-reference and extract information from these lists. These data have a "layered" structure, such that increasingly smaller subsets of the data contain more detailed information (with finer granularity). In all, the dataset records 20.2 million (20.2M) individual communications with 1.3M different donors during 2006-2011. However, most of the information pertaining to fundraiser design is available for 8.6M communications taking place during 2009-2011, and we also have more detailed information about donors for $4.3M$ of these communications. Fewer than 10% of communications result in donations, and we also have multiple layers of data for these gifts. Table 1 shows how much of each type of information is available. Below, we describe the data in more detail.

Table 1 Amount of data available for various types of information.

Type	Amount	Donors	Content	Raw size	Full size
Communications	20.2M	1.3M	Account ID, location, affiliation Campaign types of communications	1.6 GB	—
	8.6M	1.2M	All design features	1.1 GB	3.0 GB
	4.3M	531K	Donor segmentation information	609 MB	2.4 GB
Gifts	819K	366K	Dates, amounts, payment methods	105 MB	—
	308K	193K	All design features	69 MB	103 MB
	169K	98K	Donor segmentation information	39 MB	110 MB
	89K	87K	Disaster type, magnitude, location	28 MB	49 MB
	6.9K	6.5K	Segmentation+disaster information	2 MB	6 MB

"Raw size" refers to the data obtained directly from the Red Cross; "Full size" also includes additional information derived from the dataset, described in Section 5.2.

Donors. Each donor is assigned a unique *account* number, so that we can always identify the specific donor with whom any given communication occurred. The *location* of a donor is represented in our study by the U.S. state associated with an account. Limited *affiliation* information is available, e.g. whether the donor is listed with a county or city chapter. At the same time, the Red Cross does *not* have access to personally identifiable information about the donors (such as demographic or income information). This generally holds for the entire non-profit industry.

Communications with donors are classified according to *campaign type*, which reflects the status of a donor (acquisition, renewal, lapsed, etc.) at the time of the communication. Donors are further categorized according to *donor class*, a measure of how much they give per donation, which can be low ($\$10-\$99$), medium ($\$100-\$499$), or high ($\$500-\$9999$), with some additional classes such as "Haiti-influenced donors" representing connections to a particular disaster. We also have records of donor *recency*, which represents the time since their last donation (e.g. 0-6 months, 6-12 months, etc.). These two pieces of information are jointly referred to as *segmentation information.* Other donor classes may also be defined in connection with specific disasters, e.g. "Haiti-influenced donors" whose first donation followed on the Haiti earthquake.

To understand our results, it is important to bear in mind the specifics of the relationship between campaign type and donor recency. Table 2 shows the number of communications in each recency category, for five important campaign types. All donors in the Acquisition category have only made one disaster donation. As expected, many of them did so within the past six months, reflecting the effort to cultivate recent donors. However, a substantial group of communications in this category were targeted at donors who donated over 36

Table 2 Breakdown of communications by campaign type and donor recency for several important campaign types.

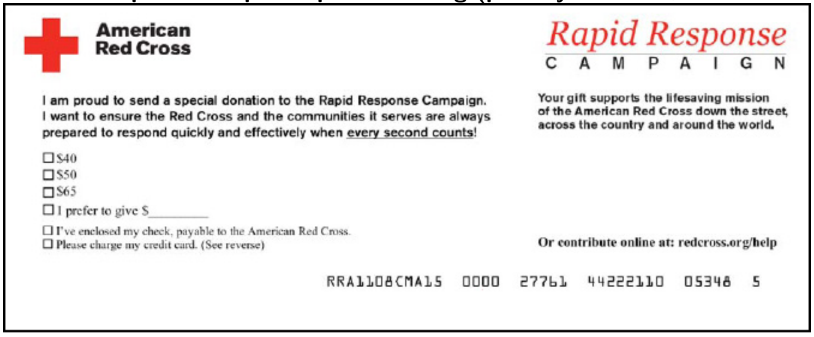
	0-6 mos.	7-12 mos.	13-18 mos.	19-24 mos.	25-36 mos.	37-48 mos.
Acquisition	52192	6313	1562	2657	0	125525
Cultivation	478665	98744	87008	3546	0	0
Lapsed	0	0	22009	17006	38158	0
Renewal	308236	407490	22366	11149	1370	770
Generic	1017854	350986	361570	79511	0	0

months ago, demonstrating a late effort by the Red Cross to reach out to donors who had never been cultivated. Among our current supporters (Renewal and Cultivation), many have made their last donation recently, but substantial proportions fall into the 7-12 and 13-18 month categories.

Donations. The *date*, *amount*, and *payment method* of every donation are recorded, as well as the *fund* receiving the donation. In the available data, 60 funds are associated with specific disasters (e.g. Iowa flood, Haiti earthquake, or Tohoku tsunami), for which we have records of *disaster type* (e.g. earthquake, flood, or hurricane), *magnitude* (death toll, cost in millions), and *location* (domestic or international).

Designs. Any given piece of mail is constructed with a set of design features. These include *personalization* of the mailing (inclusion of the donor's name and address), the presence or absence of various *donation options* (e.g. checkboxes for donating $20 or $30), and the possible inclusion of *gift items* such as mailing labels or a glowstick. A *supporter card* is included in 5.5% of communications. Cards are sent to the Renewal type, but cover all of the main donor and recency categories within that type; the gift items were sent to Acquisition and Lapsed types, again covering a variety of recency groups. A proportion of 64% of all communications offers the option to donate *online*. In 3.5% of communications, the donor has the option to *choose* the fund that will receive his or her donation. The *formulation* of the appeal is described, e.g. whether the letter mentions a specific disaster

Figure 1 Example of a rapid response mailing (publicly available; see WCAI 2012).



American
Red Cross

Rapid Response
CAMPAIGN

I am proud to send a special donation to the Rapid Response Campaign.
I want to ensure the Red Cross and the communities it serves are always prepared to respond quickly and effectively when every second counts!

Your gift supports the lifesaving mission of the American Red Cross down the street, across the country and around the world.

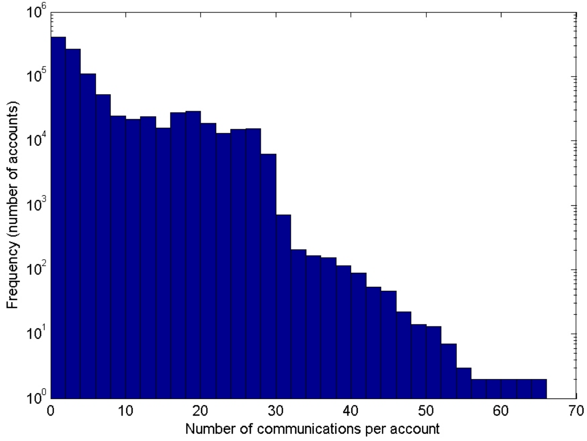
□ $40
□ $50
□ $65
□ I prefer to give $__________

□ I've enclosed my check, payable to the American Red Cross.
□ Please charge my credit card. (See reverse)

Or contribute online at: redcross.org/help

RRA1108CMA15 0000 27761 44222110 05348 5

Figure 2 **Empirical distribution (log-scale) of the number of communications per account.**



(about 40% of communications) or offers a generic story about disaster relief (10%), or whether it emphasizes disaster preparedness (50%). Figure 1 shows an example of a mailing with three suggested amounts of $40, $50 and $65. These amounts may also be *dynamically generated*, as described in Section 2.

We often have records of multiple communications with the same donor account. Figure 2 shows the empirical distribution of the number of communications per account, that is, the frequency of accounts for various numbers of communications. As expected, most accounts have fewer than 10 communications, but there are some with over 60. The decision to continue communicating with a particular donor is influenced by factors in our dataset such as the donor class and recency (high-class, recent donors are targeted more often). Note that class and recency are determined by the donor's most recent donation only; Red Cross analysts believe that this information is sufficient for deciding whether to target a donor. In general, the Red Cross also prefers to target donors who choose to give to *general* funds (such as "Where Our Need Is Greatest") rather than to specific disaster funds; however, this is not a major factor in STAART, since over 91% of all gifts in our data are made to general funds, and only 3.5% of mailings allow a choice of fund.

However, these communications may have different campaign types, reflecting the donor's transition from Acquisition to Renewal, or Renewal to Lapsed. The donor's segmentation information can also change over time. The design features can change with every communication as the organization experiments with new outreach methods. Note also that the outcome of a communication (*success* or *failure*, i.e. whether or not the communication elicited a donation) determines the *amount* of information that we receive. We can only observe detailed fund and disaster information for successful communications, where there is a record of money sent to a particular fund. Because the number of gifts is far smaller than the number of communications, the total volume of fund and disaster information is also relatively small; see Table 1 for the exact numbers. To leverage as much of the data as possible, we develop a separate model for each layer of data.

5. Methodology

We describe the methodology used to analyze the Red Cross dataset. In Section 5.1, we describe the basic statistical model that forms the foundation of our analysis. The Appendix gives the full technical details of the estimation of the model on the data. Section 5.2 discusses additional modeling and feature generation.

5.1. Statistical model and procedure

Let $i \in \{1,...,I\}$ denote the ith donor account, with I being the total number of accounts. To reflect the longitudinal nature of the data, let the *panel size* N_i be the number of recorded communications with account i, and let $N = \sum_{i=1}^{I} N_i$ be the total number of communications. Also let y_{ij} denote the result of communication $j = 1,...,N_i$ with account i.

A communication is "successful" ($y_{ij} = 1$) if it results in a donation. Otherwise, the communication is considered to be a failure ($y_{ij} = 0$). We begin with a mixed-effect logistic regression model, which assumes that

$$\mathbb{E}\left(y_{ij} \,|\, b_i\right) = g^{-1}\left(\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i\right), \tag{1}$$

where g denotes the logit link function. The p-vector $\mathbf{x}_{ij}$ represents the attributes of the jth communication with account i. This includes any relevant donor, donation and design information (see Section 4, Table 1) available for this communication. For example, a particular component x_{ijk} can be equal to 1 if the jth communication with account i included an option to donate online, and 0 otherwise.

The parameter b_i is a random effect (Laird and Ware 1982); we assume that $b_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 represents random variation between panels, and that the individual observations y_{ij} are conditionally independent given b_i. We include random effects in the model for several reasons. First, b_i can be used to represent unobservable variation in donor behaviour, specific to account i, and reduces statistical bias that arises when multiple observations come from the same source. Second, random effects reflect the fact that the donors in the dataset come from a larger population of donors, and the Red Cross continuously communicates with new individuals. Random effects thus allow us to reason about the entire population (and, potentially, new donors). Third, random effects allow for a much more compact model with only a single additional parameter σ^2, whereas adding a fixed effect for each account would add hundreds of thousands of attributes. Finally, modeling b_i as a random variable reflects the organization's considerable uncertainty about individual donor characteristics and behaviour.

For given β and σ^2, the joint probability of observing y_{ij}, $i=1,...,I$, $j=1,...,N_i$, can be written as

$$L(\boldsymbol{\beta},\sigma)=\prod_{i=1}^{I}\int_{-\infty}^{\infty}\prod_{j=1}^{N_i}\left(\frac{e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i}}{1+e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i}}\right)^{y_{ij}}\left(\frac{1}{1+e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i}}\right)^{1-y_{ij}}\frac{e^{-\frac{b_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}db_i \tag{2}$$

where the integral represents an expectation of a conditional probability given b_i. Then,

$$(\boldsymbol{\beta}^*,\sigma^*)=\arg\max_{\boldsymbol{\beta},\sigma}\log L(\boldsymbol{\beta},\sigma) \tag{3}$$

represents the maximum-likelihood estimates of $\boldsymbol{\beta}$ and σ. The MLE optimization problem in (3) is typically solved using expectation-maximization (EM) algorithms (McLachlan and Krishnan 2008) where the random effect terms are treated as missing data and estimated iteratively. However, this approach is intractable in our problem, because (2) is a product of I integrals (where I is on the order of 10^6), which cannot be expressed in closed form and must be evaluated numerically. Numerical methods such as Gaussian quadrature may be feasible for small I, N_i or p, but scale very poorly to large data (Fessler and Hero 1994, Karl et al. 2014).

All else being equal, a concise model with a smaller number of features is preferable. Hundreds of features can be extracted from the Red Cross dataset, including those described in Section 4 and the interaction terms discussed later in Section 5.2. However, from a managerial viewpoint, it is preferable to focus on a small set of key drivers of campaign success, and from a statistical viewpoint, a smaller model reduces the risk of overfitting and is easier to generalize; also, extra attributes impose additional noise on prediction. To identify the most significant features, we use model selection methodology (Hastie et al. 2001) and replace (3) by

$$(\boldsymbol{\beta}^*,\sigma^*)=\arg\min_{\boldsymbol{\beta},\sigma}\left\{-\log L(\boldsymbol{\beta},\sigma)+\lambda\sum_k|\beta_k|\right\}, \tag{4}$$

where L is as defined in (2). The tuning parameter λ represents a price or penalty incurred if we choose a non-zero value for any β_k included in the final model. The penalty function $\|\boldsymbol{\beta}\|_1 = \sum_k |\beta_k|$ is non-differentiable around zero, causing β_k to shrink exactly to zero if the kth feature is found to be insignificant. Combining MLE estimation with this type of penalty function is known as the Lasso method, proposed by Tibshirani (1996) for linear regression, with later extensions to generalized linear models (Meier et al. 2008) and mixed effects (Schelldorfer et al. 2011). Equation (4) balances the need for an accurate model with more predictive power against the need for a compact model with fewer features. By choosing λ carefully, we ensure that non-zero regression coefficients will only be assigned to attributes with a significant impact on model accuracy. The Lasso penalty can lead to substantial practical benefits in applications with $N \gg p$ (Rudin et al. 2012). See also the Appendix for additional numerical results demonstrating the benefits of this approach.

The choice of λ is automated as follows. Let $\boldsymbol{\beta}(\lambda)$ and $\sigma(\lambda)$ be the choice of $\boldsymbol{\beta}$ and σ that solves (4) for given λ. Then, let

$$\mathcal{A}(\lambda) = \{k : \beta_k(\lambda) \neq 0\}$$

be the set of attributes identified by (4) as being significant. The size $|\mathcal{A}(\lambda)|$ is the number of features included in this model. We then solve

$$\lambda^* = \arg\min_{\lambda} \left\{ -2\log L\Big(\boldsymbol{\beta}(\lambda), \sigma(\lambda)\Big) + |\mathcal{A}(\lambda)| \cdot \log N \right\}, \tag{5}$$

choosing the penalty term to minimize the well-known Bayesian Information Criterion (BIC) of Schwarz (1978). Combining BIC with Lasso is a fairly widespread technique, and has been found to yield good practical performance on a variety of problems (Wang et al. 2007). Other possible criteria from the literature include Akaike's Information Criterion (or AIC; see Akaike 1973) and cross-validation (or CV; see e.g. Breiman and Spector 1992). However, Zou et al. (2007) argues that AIC-Lasso tends to include more non-zero predictors than necessary. Furthermore, although the CV criterion is widely used in the literature, it has less theoretical support; a recent work by Homrighausen and McDonald (2013) has obtained consistency results only for a restrictive class of models with orthogonal design matrices, which rarely occur with discrete data.

Finally, we remark on the additional challenge of estimating (4) on a large dataset. Model selection reduces the feature space, but (4) remains computationally prohibitive for large N. In such circumstances, a natural strategy is to take a random sample (of a tractable size) from the data, and use this "subsample" to estimate the model. We adopt this approach; however, the statistics literature has demonstrated (Bühlmann and Yu 2002, Bradić 2014) that using a single subsample can introduce bias into the model, as well as inflate the variance of the estimated coefficients. To mitigate this issue, we draw S small subsamples, thus obtaining S distinct Lasso models. Due to the noise introduced by subsampling, the set $\mathcal{A}(\lambda^*)$ of selected features varies across subsamples. To reduce this variability and ensure that only significant features are selected, we use a "majority vote," i.e., we include the kth feature in our model if it is selected in over 50% of the S Lasso models. Please see the Appendix for the full technical details of this procedure.

Letting $\mathcal{A}^*$ be the set of all such features, we can finally recompute $\arg\max_{\boldsymbol{\beta},\sigma} \log L(\boldsymbol{\beta},\sigma)$ subject to the additional constraint that $\beta_k = 0$ for $k \notin \mathcal{A}^*$, yielding the optimal estimates of the significant regression coefficients. This step is known as "debiasing" or "post-Lasso," and has been demonstrated to eliminate bias from the Lasso estimator and produce more precise confidence intervals in settings such as least squares and quantile regression (Wright et al. 2009, Belloni and Chernozhukov 2013). Because the computational cost of this step is still prohibitive, we can perform debiasing on a new set of subsamples and average the results to obtain the final coefficients. Please see the Appendix for a detailed discussion.

5.2. Modeling and feature generation

In addition to the information already present in the data and described in Section 4, we generated additional attributes to address important statistical and modeling issues. To study the effectiveness of segment-specific fundraising strategies, we constructed numerous interaction terms and incorporated them into our model. In particular, we considered interactions between design features, such as the presence or absence of dynamic amounts, and donor features such as type (Acquisition, Renewal, etc.) and recency. This allows us to capture segment-specific effects, e.g. strategies that work better with new donors than lapsed donors. Interactions between donor features and the presence or absence of other donation options were also investigated. Model selection becomes crucial when considering interactions, as the number of two-way interaction terms grows quadratically with the number of features. The last column of Table 1 shows that the size of the data increases dramatically once the additional features have been generated.

Secondly, we generated additional control covariates to reduce bias due to unobservable correlations between communications. The bias is potentially due to correlation between measured and unmeasured, confounded, or "missing" features of each communication. Most notably, the behaviour of a single donor during the surveyed time period may be subject to time dependencies. A donor is unlikely to maintain the same level of contribution over time; rather one may expect the donor to lapse once the resources he or she has allocated for donation have been exhausted. From a behavioural standpoint, a donor who contributes frequently may simply be more motivated, or place higher value on pro-social activity. We control for the time factor as follows. For communication j with account i, we calculate 1) the number of previous communications with i, prior to the date of communication j; 2) the number of previous *successful* communications with i; 3) the number of previous communications *of the same type* with i; and 4) the number of different funds that have sent requests to the donor thus far. These attributes are included in $\mathbf{x}_{ij}$. Additional information on the time lapsed between donations is provided in the form of the recency attribute.

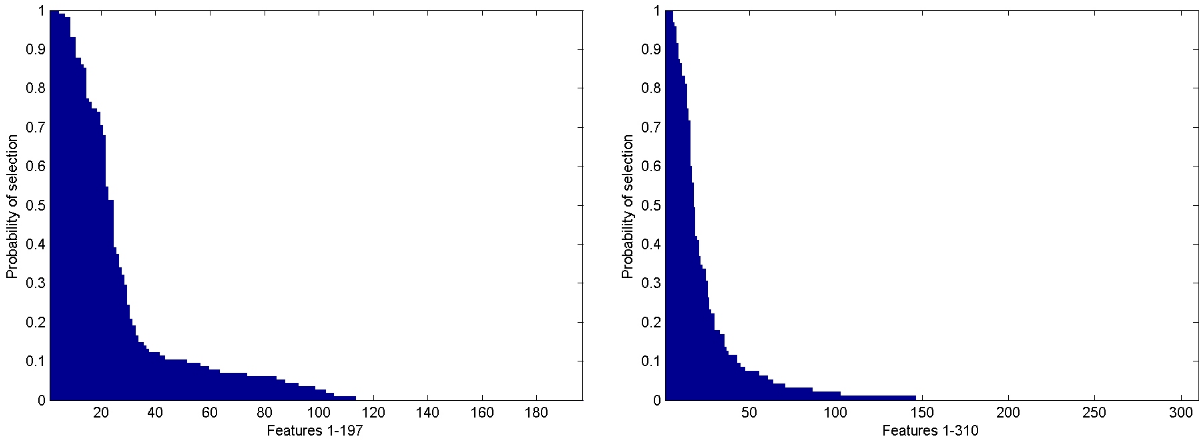
One of the research questions in Section 2 deals with the dynamic generation of ask amounts. Donor class information, based on the size of the donor's last donation, allows us to control for the magnitude of the asked amount and separate the effect of dynamic generation from the effect of the precise amount asked for. We also control for any fixed ask amounts (donation options) that appear on the mailing.

We also validated our results through cross analysis, comparing the results of feature selection across different model types. For example, we compared the results of the model in Section 5.1, where each observation corresponds to an individual communication, and a different model where an entire campaign is viewed as a single observation, and the response variable is the success rate of the campaign. While the exact values of the estimated coefficients differ between models, the key managerial implications of the results are consistent throughout the study.

6. Results

Sections 6.1 and 6.2 discuss the logistic regression model described in Section 5.1, used to predict the success probability of an individual communication. Sections 6.3 and 6.4 present additional analysis of campaigns and donations, respectively. Finally, Section 6.5 lays out simulation results illustrating the potential of our key insights to improve conversion rates.

Figure 3 **Ranking of features, in descending order of the proportion of subsamples where each feature was selected.**



(a) Base model (Section 6.1).

(b) Model with segmentation data (Section 6.2).

6.1. Communication-based models: design information

We begin with the model from (1), where y_{ij} is the outcome of the jth communication with donor i, and g is the logistic link function. Recall from Table 1 that the data have a layered structure. We apply this model to two layers. The first layer uses 8.6M communications from 2009-2011 for which design information is available; the second layer uses 4.3M communications, but adds segmentation information. We do not consider the outermost layer of 20.2M communications, because it lacks the crucial dimension of design information. We also do not consider the innermost layer of 89K communications, because fund information is only available for donations (i.e. successful communications), and is thus unsuitable for predicting the success *probability* of a communication; however, we will return to the issue of fund information in Section 6.4.

The first layer, covering 8.6M communications, gives us a total of $p = 197$ features, comprised of campaign types (binary features indicating Acquisition, Cultivation, Lapsed, etc.), design features, donor locations, and the additional covariates and interaction terms described in Section 5.2. Due to space considerations, we do not list all of these features here; rather, we list (in Table 3) the most significant features identified by the model selection procedure from Section 5.1.

We draw $S = 120$ subsamples (see the Appendix for a discussion of how S is calculated) by Monte Carlo sampling with replacement from the large dataset. We run model selection separately on each subsample, and include a feature in the final model if it is selected

Table 3 Final estimated coefficients for first layer (8.6M communications).

Rank	Feature	Avg. coefficient	Std. deviation	p-value
1	Intercept	-3.8329	0.4636	<1e-30
2	Previous successes	0.6623	0.0889	8.2531e-25
3	$50 option	0.6330	0.0990	1.3843e-14
4	Year 2009	0.7559	0.0950	<1e-30
5	$20 option/generic type	-1.9487	0.3503	1.1276e-10
6	$15 option	-0.3373	0.0681	2.5277e-8
7	Allow choice of fund	-1.8780	0.2834	6.1010e-17
8	Dynamic amt./Renewal type*	0.0810	0.0760	0.1473
9	Dynamic amt./Acquisition type	-2.1242	0.3451	1.0304e-13
10	Dynamic amt./$50 option	4.8331	1.0783	1.0715e-6
11	Dynamic amt./Lapsed type*	-3.0032	1.4797	0.0212
12	Generic story/generic type	-0.9814	0.1294	1.2341e-28
13	Supporter card	0.2881	0.0642	1.0269e-6
14	Donor city indicated	-0.1823	0.0561	4.1723e-4
15	Renewal type	0.2693	0.0891	1.1323e-3
16	Donor city/Acquisition type	0.1729	0.0651	3.8793e-3
17	Preparedness story	0.3406	0.0495	3.0188e-18
18	$50 option/Renewal type	-5.1345	1.1209	3.8085e-7
19	$30 option/$50 option	1.6980	0.2131	3.6301e-37
20	Dynamic amt./Cultivation type	-2.6663	0.6355	4.2092e-6

All features are significant at the 0.01 level except those marked with an asterisk (*).

in a sufficiently high proportion of subsamples. We view this proportion as the empirical probability of selecting a feature. Figure 4(a) shows the 197 features ranked in descending order of selection probability. We see that over 40% of these features are not selected in any subsample, suggesting that they can be safely eliminated from our model. Moreover, fewer than 15% of features are selected in over half of all subsamples.[1] Table 3 lists the top 20 ranked features, of which half are interaction terms. The technical details for the computation of coefficients and standard errors are given in the Appendix.

Managerial insights. The results provide immediate insights into the first three hypotheses in Section 2. Feature 13 shows that the presence of a supporter card exerts a positive impact on the odds of success for an individual communication (as hypothesized). Features 12 and 17 suggest that a generic story (that is, a story describing disaster relief without reference to a specific disaster) performs poorly, while a preparedness story has a positive effect. Also, no gift items (mailing labels or glowsticks) are selected in the final model, suggesting that these gifts do not exert a significant effect on the outcome. The model also does not select features representing the inclusion of a personal disaster preparedness checklist, or a photograph depicting people being helped.

[1] In this way, we obtain more conservative results by using multiple subsamples rather than just one. Aggregation reduces the risk of over-confidently reporting a feature as being significant, when in fact it may be an outlier whose apparent significance is due to spurious correlation stemming from the additional noise induced by subsampling.

The results suggest that mailings are more effective when they focus on disaster *preparedness*, rather than on post-disaster relief efforts. This result is somewhat surprising, as it runs counter to the conventional wisdom (even among some prospective donors, surveyed by Breeze and Dean 2012) that more visceral, "emotive" imagery translates to more donations. As discussed in Section 2, empirical studies such as Bennett and Kottasz (2000) and Bennett (2009) have also found evidence that emotive content is more likely to elicit donations.

One possible explanation is that donor *cultivation* programs, such as STAART, are targeted at a specific audience, whose characteristics differ from that of the broader pool of prospective donors. In this light, an interesting connection can be made to a study of hospice donations by Bennett (2009), which found evidence that emotively designed webpages led to a higher overall volume of online donations. However, among 239 donors who submitted in-depth information about their motivations, the authors also identified a group of 101 donors, who had a history of giving to charity, and who also reacted better to informative rather than emotional content. Similar results were observed by Diamond and Gooding-Williams (2002) for donations to a homeless shelter.

The donors in the STAART dataset have all made at least one donation prior to their inclusion in the program, and thus all have at least some experience making donations. Unsurprisingly, we also see that a longer history of giving (represented by feature 2) contributes positively to the outcome. However, more informative, preparedness-oriented content contributes an additional positive effect, while more visceral stories and even visual images of people being helped appear to have no impact, or even a negative one. Additionally, a supporter card emphasizing the donor's identity as a conscientious supporter also contributes positively. Thus, we do not argue that preparedness-oriented appeals will attract more donations in every setting; rather, we argue that such appeals are more effective in the specific setting of cultivating and retaining donors.[2]

The results also provide some initial insight into the last question in Section 2. Recall that dynamic generation of ask amounts occurs in 44% of all communications, making it a significant component of the organization's strategy. Table 3 suggests that the impact

[2] There is an interesting question here whether preparedness stories can influence donors to become more committed givers, or whether they simply are better at reaching donors who were already predisposed toward informative content. This question is outside the scope of our paper; however, the managerial implication here is that preparedness stories, and informative content in general, play a significant role in cultivation efforts.

Table 4 Final estimated coefficients for second layer (4.3M communications).

Rank	Feature	Avg. coefficient	Std. deviation	p-value
1	Intercept	-3.2204	0.0336	<1e-30
2	Previous successes	0.1298	0.0065	6.6399e-29
3	Year 2010	-0.4265	0.0253	6.6727e-25
4	\$15 option/\$20 option	-2.0118	0.1816	1.5356e-16
5	0-6 mos. recency/Low donor class*	0.0814	0.0398	0.0226
6	Supporter card	0.5744	0.0449	3.0471e-19
7	Generic story	-0.7580	0.0519	6.9461e-22
8	Haiti-influenced donors	-0.3055	0.0389	3.9877e-11
9	Dynamic amt./0-6 mos. recency	0.1100	0.0307	3.4019e-4
10	Acquisition type/0-6 mos. recency	0.7042	0.1317	7.1230e-7
11	Preparedness story	0.4060	0.0359	6.5999e-17
12	37-48 mos. recency	-0.2995	0.1124	4.9178e-3
13	Specific disaster story	-0.5899	0.0336	7.9810e-26
14	13-18 mos. recency	-0.2742	0.0498	3.8371e-7
15	Allow choice of fund/0-6 mos. recency*	0.0878	0.1579	0.2902
16	Dynamic amt./Renewal type	-0.1774	0.0535	7.7752e-4
17	Renewal type/Low donor class	0.2041	0.0543	1.9305e-4

All features are significant at the 0.01 level except those marked with an asterisk (*).

of this strategy is dependent on the campaign type. The effect appears to be positive for the Renewal type, representing current supporters of the program, and negative for the Acquisition and Lapsed types, representing new and lapsed donors, respectively. However, features 8 and 11 have high standard errors. Furthermore, Table 2 shows that donors in these types exhibit substantial heterogeneity with regard to their recency. We examine this issue in more detail in Section 6.2, where we consider a smaller but richer layer of the dataset.[3]

6.2. Communication-based models: design and donor information

The next layer of data uses 4.3M communications, but adds segmentation information in the form of two groups of features representing donor class and recency. After adding interaction terms between these new features and the design information available from before, the total number of possible features in our model is $p = 310$. However, Figure 4(b) shows that, once again, only a small number of these features is consistently identified as significant. Indeed, only 17 features were selected in at least 50% of subsamples; these are listed in Table 4 with aggregated estimates, standard errors, and p-values.

[3] One interesting observation that does carry over to later models is that no location-related feature (e.g. the state where a donor lives) was ever selected, suggesting that it is more important to find a single good design for a campaign than it is to selectively target different regions with different campaigns. To our knowledge, the Red Cross does not engage in this kind of regional targeting.

Managerial insights. Most importantly, Table 4 corroborates our previous findings regarding the first three hypotheses from Section 2. Both supporter cards (feature 6) and preparedness-oriented stories (feature 11) carry significant positive effects. Unsurprisingly, a specific disaster story works better than a generic disaster story (features 7 and 13). What is more surprising (but in line with our interpretation from before) is that both specific and generic stories carry negative effects. Additionally, a special class of donors whose first contribution was influenced by the Haiti disaster (feature 8) exhibits a significant negative effect. We note that the Haiti disaster received a great deal of media attention, and thus this donor pool may contain more impulsive donors who value emotive over informative content, and may be unlikely to convert into regular supporters.

Dynamic amounts present a more complex issue. We see that this strategy now appears to produce a negative effect when applied to the Renewal type (feature 16), which seems to contradict the findings of Table 3. At the same time, the same strategy produces a positive effect for the “0-6 mos. recency” category, which contains donors from multiple campaign types (Table 2). Feature 10 also suggests interactions between campaign type and donor recency. To clarify this issue, we ran a version of the model in which features 9, 10, and 16 were replaced with three-way interaction terms. The estimated coefficients for these features are given in Table 5. For the other features in the model (carried over from Table 4), the estimated coefficients changed slightly in magnitude, but kept the same signs as before, so we omit them out of space considerations.

The results indicate that dynamic amounts are effective for recent donors (0-6 mos. recency) who have not yet converted (Acquisition type). Earlier, in Table 3, the strategy appeared to work poorly on the Acquisition type; however, Table 2 shows that 2/3 of the communications in this type actually targeted donors with a very long recency (37-48 mos.). From Table 4, we see that these donors, unsurprisingly, do not respond, leading to an overall negative effect on the Acquisition type. However, if we consider only those

Table 5 Final estimated coefficients and standard deviations for three-way interactions.

Feature	Avg. coefficient	Std. deviation	p-value
Dynamic amt./Renewal type/0-6 mos. recency*	-0.0880	0.0543	0.0549
Dynamic amt./Renewal type/7-12 mos. recency	-0.2177	0.0645	6.4953e-4
Dynamic amt./Acquisition type/0-6 mos. recency	0.7822	0.1284	4.0931e-8
Dynamic amt./Generic type/0-6 mos. recency*	0.0463	0.0370	0.1079

Features are significant at the 0.01 level unless marked with an asterisk (*).

unconverted donors whose first donation was made in the past six months, we see that dynamic amounts have a significant positive effect.

By contrast, the strategy exhibits a negative effect for the Renewal type, particularly on less recent donors (7-12 month recency). For recent new donors, who have made their first disaster donation within the past six months, there is an opportunity to "strike while the iron is hot" by offering them the chance to replicate their behaviour, this time in the role of "Red Cross supporter" rather than "disaster donor." However, for donors who have made a second donation and already converted into the STAART program, this approach is no longer effective.

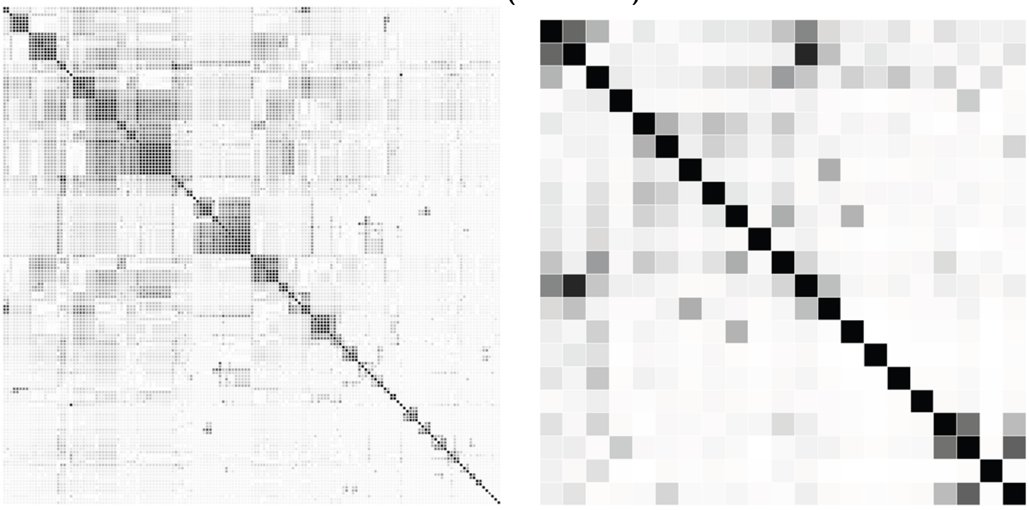
We note that all dynamic amounts in use by the Red Cross use a scale that includes, or is close to, 100% of a donor's most recent donation. We do not argue against all possible dynamic scales. However, the evidence suggests that it is unrealistic to use dynamic amounts that essentially ask a donor to maintain the same donation amount in the long term. It may be useful to experiment with other scales that, for example, use much lower ask amounts for longer donation histories. The main managerial implication of our results is that the best chance to influence donors to repeatedly give the same amount occurs at the very beginning of their donation history, before conversion occurs.

Finally, we briefly note that both Tables 3 and 4 indicate that, all else being equal, there were fewer donations in 2010 than 2009. This is particularly clear in Table 4, where "Year 2010" has a strong negative effect, while no other year is even selected. This is noteworthy, since the stock market performed especially poorly in 2009. List and Peysakhovich (2011) found a positive correlation between stock market performance and charitable donations, but observed that the effect appears to be lagged; under this model, poor stock market performance in 2009 would be expected to lead to fewer donations in 2010, providing a possible explanation for the pronounced negative effect of 2010 in Table 4.

6.3. Campaign-based models

We constructed a different set of models that considered the data in Sections 6.1 and 6.2 from another viewpoint. These models aggregate the set of communications by campaign. Again, we use the logistic model in (1), but now y_{ij} represents the success rate of campaign i on the jth donor segment. Success rate is expressed as the ratio of the number of successful communications (resulting in donations) to the total number of communications in the campaign. For example, if the organization mailed 100 copies of a letter to a certain class of

Figure 4 Empirical correlations of (a) all features and (b) selected features. Darker colours represent heavier correlation (closer to 1).



(a) Heatmap of 157 features.

(b) Heatmap of 21 selected features.

donors, and received 6 gifts in response, the success rate of the letter is 0.06 for that donor segment. A simple modification of (1) allows us to consider continuous-valued observations between 0 and 1. The main purpose of this analysis is to corroborate the results obtained in Sections 6.1 and 6.2 and demonstrate that similar results emerge without the subsampling techniques described in the Appendix. Our discussion focuses on the second model (Section 6.2, segmentation information included); we show that the relationships observed in this model also hold when the data are reorganized for campaign-centric analysis.

The attributes $\mathbf{x}_{ij}$ of the ith campaign and jth segment are largely the same as in Sections 6.1 and 6.2, and interaction terms are constructed as in Section 5.2. However, we are not able to include features that are not in one-to-one correspondence with campaign segments. For example, donor location varies on the level of individual communications, as the same letter can be mailed to people in different states. With recency and donor class included in the model, the total number of features was $p = 157$, with $I = 60$ panels (campaigns) and $N = 952$ campaign segments in all. The size N_i of each campaign ranges up to 132 segments.

The relatively small size of this dataset allows for a tractable analysis on its entirety, without the need for small subsamples. At the same time, aggregation across campaigns leads to the new problem of inflated empirical correlation, stemming from the relatively small magnitude of N relative to p. Figure 5(a) shows empirical correlations between all

Table 6 Final estimated coefficients for the campaign-based model with segmentation information.

Rank	Feature	Estimate	Std. error	p-value
1	(Intercept)	-3.12390	0.07434	<1e-30
2	\$15 option/\$20 option	-1.40281	0.12573	3.1995e-29
3	0-6 mos. recency/Low donor class	0.55805	0.07598	1.0679e-13
4	13-18 mos. recency	-0.40337	0.06540	3.4145e-10
5	7-12 mos. recency/Low donor class	0.50571	0.08343	6.8060e-10
6	Specific disaster story	-0.38501	0.07145	3.5228e-08
7	Card	0.66038	0.18892	2.3262e-04
8	Includes specific fund/37+ mos. recency	-0.79300	0.26354	1.3062e-03
9	Generic story	-0.33655	0.13033	4.9400e-03
10	Preparedness story	0.21504	0.08808	7.3436e-03
11	Dynamic amt./0-6 mos. recency	0.22205	0.10655	0.0187
12	Dynamic amt./Lapsed type	-0.86614	0.43200	0.0227
13	Allow choice of fund/0-6 mos. recency	0.61133	0.33262	0.0328
14	0-6 mos. recency/Haiti-influenced donors	0.18883	0.11368	0.0484
15	Followup type	0.25435	0.17342	0.0707
16	Includes specific fund/7-12 mos. recency	-0.50035	0.44334	0.1292
17	0-6 mos. recency/High donor class	0.17362	0.16390	0.1445
18	Dynamic amt./Renewal type	-0.18834	0.20609	0.1814
19	19-24 mos. recency/Haiti-influenced donors	0.15064	0.21575	0.2419
20	Renewal type/Low donor class	-0.13402	0.19576	0.2482
21	Option to donate online/High donor class	0.11607	0.19923	0.2809
22	Renewal type	-0.01884	0.15227	0.4522

"Std. error" refers to the usual statistical standard error of an estimated coefficient. The first 14 features were significant at the 0.05 level.

157 features; in particular, the dark blocks visible in Figure 5(a) show that, for certain groups of features, the empirical correlations are close to 1. This issue complicates statistical analysis, as the matrix $\mathbf{x}^T\mathbf{x}$ is not invertible. Random effect models are even more sensitive to correlation, as the number of panels is even smaller than the sample size. Model selection solves this issue by reducing the number of features from 157 to just 21, not counting the intercept. Figure 5(b) shows that these selected features exhibit much lighter correlation. Table 6 shows the final results for this model, ranking the selected features by p-value.

Managerial insights. It is most relevant to compare Tables 4 and 6, because they both include segmentation information. With this in mind, we see that Table 6 reproduces our key findings from Section 6.2. Most crucially, we observe identical insights on relief vs. preparedness: both generic (feature 9) and specific (feature 6) disaster stories carry negative effects. By contrast, preparedness-oriented campaigns (feature 10) have significantly higher success rates. Furthermore, supporter cards (feature 7) continue to contribute to campaign success.

We also considered a campaign-centric version of the model from Section 6.1 (design information, but no segmentation). We do not give the full details here for space considerations, as they mostly repeat our previous discussion. However, we briefly note that this model produced the same results with regard to preparedness vs. relief, as well as supporter cards.

6.4. Gift-based models

We also considered another set of models incorporating donation and disaster information. That is, we link an incoming donation to a specific disaster with attributes such as type (flood, earthquake, etc.), location (foreign or domestic), and magnitude (e.g. death toll). Each donation is associated with a monetary amount and a payment method, which we have also not discussed up to this point. We mostly obtain the negative result that disaster-specific information does not significantly affect donations, and include the discussion below for completeness.

Donation and disaster information is difficult to include in our previous models, because it can only be observed for successful communications (those that result in donations). Russ Reid has confirmed that there is no way to connect an unsuccessful communication to a specific fund. While the literature offers models for handling "donations" of size zero, our situation is more complicated because we also observe additional information (extra features) when a communication succeeds that is not observable if the communication fails. We chose to conduct a separate analysis that is confined to gifts only, removing all unsuccessful communications. That is, we return to equation (1), but now define y_{ij} to be the dollar amount of the jth *gift* contributed by the ith account. The link g is chosen to be the identity function, corresponding to a linear regression model.

Fewer than 5% of all communications are successful, which drastically reduces the model size. Table 7 shows the sizes of models fit to different layers of data; no model is large enough to require subsampling. For space considerations, we do not list the full results from all four models here, but we highlight the main points from all four models. Note that, due to the smaller size of these models, fewer features are statistically significant at the 0.05 level. The last column of Table 7 counts these features for each layer.

Disaster attributes appeared to have no significant impact on donation amounts. Of the 14 features in Layer 4 that were statistically significant at the 0.05 level, only one involved a disaster attribute. This was the interaction "Earthquake/High donor class." It is unsurprising that the high donor class should have a strong positive correlation with donation amount, as this class includes the largest gifts, up to $9999. As for the earthquake

Table 7 Sizes of gift-oriented models.

Layer	Size	Features	Interactions	Total	No. selected	0.05 level
1: All gifts, 2009-2011	309,451	96	33	129	52	34
2: Segmentation only	168,588	104	163	267	68	51
3: Disaster only	89,529	103	104	207	23	17
4: Segmentation+disaster	6,908	108	228	336	23	14

attribute, we note that it applies to the Haiti disaster, which was widely publicized and led to a high volume of donations.

Of the 34 statistically significant features in Layer 1, 18 corresponded to different donor locations (represented by U.S. state). Potentially, this suggests that there may be regional differences in donation amounts (though not in campaign success rates; see Sections 6.1-6.3). However, as segmentation and disaster information was added in Layers 2-4, the number of statistically significant locations shrunk to 6/51 in Layer 2, 9/17 in Layer 3, and just 2/14 in Layer 4. On this basis, we argue that the impact of campaign design and donor segmentation is much greater, and more important for policy decisions, than the possible impact of regional differences.

Design attributes seemed to have relatively little impact on donation amounts. Layer 4 contains only a single significant feature involving such an attribute, the interaction "Dynamic amt./High donor class," with a significant positive effect. However, there are several significant features involving the high donor class, all with strong positive effects, suggesting that the effect is more likely due to the fact that donors in the high donor class simply give more money to begin with.

Managerial insights. Most of the statistically significant features selected in Layers 1-4 were related to donor attributes such as recency and low/medium/high class, in contrast with our results from Sections 6.1-6.3. This suggests that a well-designed appeal may get more donors to respond (increasing the campaign success rate or the probability of receiving a donation from a particular donor), but the *amounts* of their donations are largely determined by immanent donor characteristics such as the donor class. Of course, this result should be considered in the specific context of *cultivation* campaigns.

6.5. Simulation results

We conducted a simulation study to quantify the potential benefits of the insights in Sections 6.1-6.4. By simulating donors, we can compare historical fundraising strategies with our recommended ones. It is difficult to evaluate our recommendations based purely on the historical data, since the data represent the actual outcomes of a particular set of design choices used in the past, and there is no way to redo those same communications with a different set of designs.

For our simulations, we randomly sampled 10,000 donors who received mailings during the first six months of 2009. For the ith donor account, we randomly generate a value $\hat{b}_i$ from the random effect distribution estimated in our statistical analysis. These values are viewed as fixed in the subsequent procedure. Then, for the jth historical communication with account i within the six-month period, we simulate an outcome $\hat{y}_{ij}$ from a Bernoulli distribution satisfying $\mathbb{E}\left(\hat{y}_{ij}\right) = g^{-1}\left(\mathbf{x}_{ij}^T\bar{\boldsymbol{\beta}} + \hat{b}_i\right)$, where g is the logit link function and $\bar{\boldsymbol{\beta}}$ is a vector of the final estimated coefficients from Section 6.2 (including the three-way terms).

The vector $\mathbf{x}_{ij}$ can now be modified to reflect different fundraising strategies. The *historical strategy* simply consists of setting the elements of $\mathbf{x}_{ij}$ equal to their historical values. The *new strategy* uses the following rules. First, a supporter card is always included in the first communication with donor i that is of the Renewal type, but not in any other communications with that donor. Second, dynamic amounts are only applied to new, unconverted donors, as discussed in Section 6.2. Third, generic stories are never used; preparedness and specific stories are each used 50% of the time (we assumed that the organization may prefer to use a variety of stories, even if one type works better than another). Fourth, gift items are never included. For the first communication with a donor, descriptive features such as recency are always set to their historical values.

In this way, we can generate donors with realistic features, as well as outcomes for two versions of the same communication that have different design features. We updated time-dependent features dynamically for both strategies. For example, for the ith donor, we store a counter representing the number of successful communications with that donor. The counter is incremented if $\hat{y}_{ij} = 1$ (for the particular strategy used) and used to set the "previous successes" feature for the next communication with the donor. Likewise, if $\hat{y}_{ij} = 1$, the recency of donor i is reset to "0-6 months" for the next communication.

On average, the sampled donors receive 12-13K mailings in the six-month period. The 99% confidence intervals for the success rates achieved by the two strategies are 0.0539 ± 0.0006 for the historical strategy, and 0.0812 ± 0.0009 for the new strategy. The new strategy significantly improves the success rate. Additionally, when $\hat{y}_{ij} = 1$, we can plug $\mathbf{x}_{ij}$ into the models in Section 6.4 to obtain predicted donation amounts for the simulated successes (this also allows us to dynamically update the donor class features for the next communication with donor i). On average, the historical strategy collects \$61,231 ± \$1,167 in revenues, while the new strategy collects \$99,559 ± \$1,453 (99% confidence intervals reported for both strategies).

We should note several grounds for caution in interpreting this comparison. First, all success probabilities are calculated from the estimated model, although this model was calibrated using a massive volume of historical data. Second, the simulations assume that, in both scenarios, the Red Cross sent the same number of mailings to the same donors. However, this may actually cause the simulations to under-report the improvement achieved by the new strategy: in practice, if the organization were to receive more donations, it would also tend to communicate with those donors more frequently, thus creating an opportunity for still more successes. Ultimately, while the precise numerical improvement achieved by the new strategy reflects the assumptions made in our simulations, these results suggest that the managerial insights from Section 6 can be translated into a few simple design rules that offer significant potential for improving retention rates, even under a pre-specified number of communications with each donor.

7. Conclusion

This paper presents a data-driven study of disaster donor cultivation, using a massive dataset from the American Red Cross, to formulate several models of fundraising success. The results of this analysis lead to the following managerial insights for managers at the Red Cross and other non-profits who work on *cultivation* of first-time donors into regular supporters:

1. *Relief vs. preparedness.* The influence of emotive imagery on charitable behaviour is widely recognized. However, in donor *cultivation*, there is evidence to suggest that preparedness-based appeals are much more effective than relief-based appeals. More broadly, this suggests that non-profits may benefit from more informative (rather than emotive) content in programs that focus on cultivation.

2. *Supporter cards.* A small card affirming a donor's identity as a Red Cross supporter appears to exert a significant positive impact on cultivation efforts. We recommend the inclusion of this item as a standard component of STAART mailings, perhaps in the first communication with a donor. We believe that non-profits in general can improve donor retention by using such techniques to reinforce the identity of potential supporters.

3. *Gift items.* On the other hand, other gift items, such as emergency lights and address labels, appear to have no significant effect on cultivation. From the evidence, we conclude that these items can be eliminated as a cost-saving measure.
4. *Dynamic amounts.* The strategy of dynamic amount generation individualizes the ask amounts in a mailed appeal, based on each donor's previous donation history. Essentially this strategy encourages a donor to maintain an earlier level of contribution. The evidence suggests that this works on very recent first-time donors who have not yet been converted, but may actually be counterproductive with current and lapsed supporters. In such cases, it may be better to use a few standard ask amounts, or substantially scale down the dynamic amounts.

A major challenge of donor cultivation, and a limiting factor of this study, is the relative lack of information on donors. Previous work on donor behaviour has drawn from surveys and policy studies, which cover a relatively small number of individuals, but provide detailed information on income, demographics, donor motivation, and other relevant factors. At the same time, while these attributes are valuable in understanding the economic and behavioural drivers of donations, they are unobservable to organizations like the Red Cross during *operational* decisions. We have formulated recommendations to help non-profit managers to improve cultivation campaigns based on information that they have available at the time the decision is made. To our knowledge, this is the first paper to adopt a data-driven approach to this problem. We believe this to be an important contribution to the study of non-profit donations.

Appendix. The challenge of massive data

From the point of view of traditional statistics, the logistic regression model described in Section 5.1 should always benefit from more data. From a purely theoretical viewpoint, a large sample size N is always a good thing, and theoretical issues arise only when $N < p$. However, in practice, the *estimation* of the model becomes computationally intractable when N is in the millions. We emphasize that the computational challenge arises, not from memory issues (various techniques and software packages, e.g. `biglm` in R, can be used to address that issue), but rather from the estimation of (1), which requires us to optimize an expensive, highly non-linear function.

Recall that (2) is a product of I integrals, where I is the number of unique donor accounts (over 1M in all). Furthermore, each integrand is a product of N_i logistic functions with a normal density, and thus is highly non-linear and non-convex. None of the I integrals has a closed-form solution; consequently, (2) can only be evaluated numerically, e.g. using Monte Carlo integration or Gaussian quadrature. Numerical integration introduces additional error into the evaluation of the likelihood function, and is also expensive for large I and N_i since each integrand must be evaluated multiple times. For these reasons, quadrature methods are infeasible for large problems, leading to both memory and convergence issues for expectation-maximization (EM) algorithms. This issue is well-known in the literature; for example, Karl et al. (2014) finds that EM algorithms scale poorly to large datasets. In our experience, the available computational procedures for solving (3) with random effects simply stalled, crashed, or otherwise failed to produce meaningful results.

With the advent of increasingly large datasets, the statistics literature has now begun to pay closer attention to large-sample data, where N is very large (in the millions) and p is moderately large (several hundred). Even with a large number of samples, such data may be vulnerable to noise accumulation, spurious correlations, and algorithmic instability (Fan et al. 2014). Ideally, statistical methods for such data should be computationally tractable while retaining the theoretical guarantees of classical statistics (such as consistency). In order to scale up to the Red Cross dataset, we synthesize several emerging statistical methodologies, such as small-sample bootstrapping and stability selection, that yield both tractable and rigorous results.

Our approach is based on the idea of "subsampling," or conducting the statistical analysis on a small, randomly generated subset of the large dataset. This is a natural strategy for dealing with big data, since a small subsample remains, in some sense, representative of the data as a whole. The size M of the subsample can be less than 1% of the total sample size N, allowing us to perform model selection and estimate the mixed-effect model within 1-2 hours.[4] However, if only a single subsample is used, several pitfalls may arise: 1) The statistics literature demonstrates (Bühlmann and Yu 2002) that using a single smaller-order subsample can bias the outcome of model selection by introducing false positives. This can occur if $M < \frac{N}{5}$, which is certainly the case in our application. 2) Furthermore, Bradić (2014) proves the stronger result that, if only a single subsample is considered, it is virtually impossible to retrieve a sparse set of significant features (that is, the probability of doing so vanishes to zero). 3) Subsampling introduces additional noise into the problem. Thus, a single subsample may inflate the variance of the estimated coefficients, analogous to how the variance of a classical sample mean is larger when the sample size is smaller. 4) A feature that appears infrequently in the big data (e.g., the supporter card feature) may be misrepresented in the subsample. Multiple subsamples can give a more representative picture.

To mitigate these issues, we draw S small subsamples, leading to S distinct, independently estimated Lasso models. Each subsample will produce different results: the number of selected features may vary across subsamples, and the set of selected features itself may vary. However, as we describe below, these results can be aggregated to obtain a single final set of accepted features, regression coefficients, and standard errors. Recent work in statistics (Kleiner et al. 2012, 2014, Bradić 2014) proves that, if M and S are correctly chosen, the aggregated results retain theoretical properties such as consistency, and correct bias that may arise with a single subsample.

[4] The cost of estimation is superlinear in the problem size; consequently, if the sample size is reduced by a factor of 100, the computational savings are much greater.

We separate the estimation procedure into two stages: we perform variable selection first, removing insignificant features to produce a model of reduced size, and then estimate random effects to correct for variation between donors. See e.g. Fan and Li (2012) for a theoretical treatment of an approach separating fixed effect and random effect estimation. Both stages use subsampling to address big data issues.

Model selection. We perform subsampling in line with the technique of Kleiner et al. (2012, 2014) as follows. For each of S subsamples, we draw M communications with replacement from the complete dataset. The work by Kleiner et al. (2014) recommends setting $M = N^{\gamma}$ for $\frac{1}{2} \leq \gamma < 1$, and obtains robust empirical results for $\gamma = 0.7$. For a dataset with $N = 8.6 \times 10^6$ communications, the size of a single subsample is $M \approx 71,500$. With regard to the number of samples, a common technique (Hastie et al. 2001) is to use $S = \frac{N}{M}$, or approximately $S \approx 120$.

We then perform model selection as in Section 5.1, replacing N by M in (5); however, as long as $M > p$, BIC preserves its theoretical consistency properties, which means that it will still correctly identify significant features (Zhang et al. 2010). To aggregate the results, we use a version of the stability selection criterion of Meinshausen and Bühlmann (2010) as follows. Each subsample $s = 1, ..., S$ produces a different solution λ_s^* of (5), and a different acceptance set $\mathcal{A}(\lambda_s^*)$. Intuitively, the kth feature is more likely to be significant if it is selected by a larger number of these subsets. We include the kth feature in our final model if

$$\frac{1}{S}\sum_{s=1}^{S} 1_{\left\{k \in \mathcal{A}(\lambda_s^*)\right\}} \geq \rho, \tag{6}$$

that is, the proportion of samples in which k is selected exceeds a threshold $\frac{1}{2} < \rho < 1$. Note that the extreme cases $\rho = 0$ and $\rho = 1$ correspond to the union and intersection, respectively, of the sets $\mathcal{A}(\lambda_s^*)$. Let $\mathcal{A}^*$ be the set of all k for which (6) holds.

Estimation. To correct for unobserved variation between donors, it is necessary to refit the random effects model of (1) with the additional constraint that $\beta_k = 0$ for $k \notin \mathcal{A}^*$ (as proved in Belloni and Chernozhukov 2013, this also corrects bias in the regression coefficients). However, even with this reduction in the size of the model, (3) remains prohibitively expensive to compute for the entire dataset. Again, we approach this problem through subsampling. To preserve the longitudinal structure of the large dataset across all subsamples, we now use entire panels as the unit of sampling. We modify the BLB technique to include importance sampling from the empirical distribution of the number of communications per panel (shown in Figure 2).

Formally, this is done as follows. Let $M' = I^{\gamma}$ be the number of donors included in each subsample, and let $S' = \frac{I}{M'}$ be the number of subsamples generated. A single subsample is created by simulating M' realizations of a discrete random variable Z with pmf

$$P(Z = i) = \frac{N_i}{\sum_{i'=1}^{I} N_{i'}}.$$

Let $Z_1, ..., Z_{M'}$ denote these M' sampled values. For each $m' = 1, ..., M'$, if $Z_{m'} = i$, we add N_i communications $y_{i,1}, ..., y_{i,N_i}$ to the subsample. In this way, a particular panel has a higher probability of being sampled if it contains more communications. Furthermore, if a panel is sampled, we automatically add every communication in that panel to the subsample, thus preserving the longitudinal nature of the data.

It remains to obtain a single set of estimated coefficients from the results of subsampling. We reoptimize (3), subject to $\beta_k = 0$ for $k \notin \mathcal{A}^*$, independently on each of the S' new subsamples. Let $\hat{\beta}_{k,s'}$ be the estimated coefficient of feature k returned by (1) on subsample $s' \in \{1, ..., S'\}$. We calculate

$$\bar{\beta}_k = \frac{1}{S'} \sum_{s'=1}^{S'} \hat{\beta}_{k,s'}$$

and report this as our final estimate of the effect of feature k. In words, we aggregate the results of subsampling by simply averaging the estimated coefficients across subsamples. Under available consistency results for subsampling, this average should converge to the true coefficient β_k with enough subsamples. Then, we let

$$\hat{\sigma}_k^2 = \frac{1}{S'-1} \sum_{s'=1}^{S'} \left(\hat{\beta}_{k,s'} - \bar{\beta}_k \right)^2 \tag{7}$$

be the sample standard error of the regression output across subsamples. We then use $\frac{\bar{\beta}_k}{\hat{\sigma}_k}$ as the relevant t-statistic, with $S'-1$ degrees of freedom, for the null hypothesis that $\beta_k = 0$. Standard techniques can be used to calculate a p-value.

We briefly discuss the choice of (7) to calculate standard errors. Notice that (7) is calculated based only on the estimated coefficients in our subsamples, not on the estimated standard errors produced by the regression model within each subsample. A recent work by Efron (2014) has argued that these within-subsample standard errors do not contribute to the asymptotic standard error of the aggregated estimator $\bar{\beta}_k$. Moreover, Efron (2014) argues that, in fact, (7) over-estimates the true variance, meaning that $\hat{\sigma}_k^2$ will produce conservative confidence intervals. For our purposes, this conservative estimator is sufficient to evaluate the significance of our results.

To summarize, we analyze the massive Red Cross dataset by separating statistical estimation into two stages. The first stage selects the most important features by conducting Lasso-type regularization on each bootstrapped subsample, then aggregating the results with stability selection. The second stage removes all features $k \notin \mathcal{A}^*$, and corrects for the unobserved variation between donors by estimating random effects in this reduced model. In addition to the theoretical advantages of aggregation, we can see from Figures 4(a)-4(b) that our approach empirically produces more conservative feature sets – there is clearly a small core of features that are "agreed" on by a majority of subsamples, but there are also clear outliers in the "tails" of the histograms that are selected in a very small proportion of subsamples (or in just one subsample).

Numerical illustration. We briefly illustrate the advantages of GLMM-Lasso with subsampling over a rougher but simpler technique, namely ordinary logistic regression (LR), in terms of two standard performance metrics (see, e.g., Smithson and Merkle 2014 for details). We compare these methods using 5-fold cross-validation (CV), a common technique in data mining for evaluating the predictive power of a model. First, we compare the deviance residuals achieved by both methods (averaged over the 5 folds in CV). The comparisons are carried out individually on 10 different subsamples, each of size N^γ. (As we discussed earlier, it is always necessary to run models on small subsamples in order to tractably obtain results.) The logistic regression model does not perform any model selection; thus, the results illustrate the benefits of using a more parsimonious model with fewer features.

Table 8 Deviance residuals of GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	100.2518	12.30325753
2	41.28834	12.61857265
3	57.07833	12.42619023
4	74.61048	12.8426679
5	86.99473	12.73456152
6	73.44249	12.28891669
7	145.982	12.39356766
8	53.33318	12.2563628
9	79.01672	12.31632096
10	30.35353	12.27578767

Table 9 Area under the ROC curve for GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	0.51984	0.70489
2	0.52615	0.70022
3	0.50530	0.69283
4	0.51175	0.70806
5	0.53812	0.69613
6	0.51591	0.70018
7	0.53985	0.71100
8	0.52010	0.68855
9	0.52511	0.69241
10	0.52399	0.69920

Table 8 presents the results of this comparison. Our model outperforms LR (achieves lower deviance) in each subsample. The results are also much more consistent for the aggregated Lasso model (LR fluctuates more across subsamples), suggesting that there is significant benefit in aggregating over multiple subsamples to reduce variance. Recall also from Figures 4(a)-4(b) that aggregation leads to more conservative results: by eliminating outlier features that are not selected by a majority of subsamples, we reduce the risk of over-confidently reporting significance.

Next, we compare the area under the ROC curve for both methods. This metric is widely used as a measure of accuracy when the data has binary responses with a small proportion of 1s (as is the case in our application). Results for 10 subsamples are given in Table 9. The Lasso model consistently outperforms LR (achieves higher AUC). Furthermore, LR generally has poor predictive power (AUC close to 0.5).

These results are quite consistent with what is known about Lasso in the literature. Classical models, such as logistic regression, estimate coefficients for a large number of features that Lasso simply removes from the model. Consequently, any prediction made by such models is subject to a much higher level of noise. Even if a plain LR model were to accurately estimate some of the coefficients, these accurate estimates are essentially drowned out by a large number of inaccurate estimates for other features. This issue, known as "noise accumulation," is quite common; for example, Fan et al. (2014) discusses how the performance of LR is often no better than random guessing in the presence of noisy data. Furthermore, simple linear models may produce over-inflated standard errors when the data is subject to a high degree of empirical correlation, an issue discussed in Section 6.3. In such settings, the p-values produced by LR may themselves be unreliable (Schaefer 1986), while Lasso is known to be less vulnerable to this issue. These examples illustrate the benefits offered by model selection in analyzing large datasets.

Acknowledgments

The authors gratefully acknowledge the help of Aleksandar Bradić with data processing for this study. We are also grateful to the American Red Cross, Russ Reid Company, and the Wharton Customer Analytics Initiative for providing the opportunity to work with the dataset that forms the basis for this paper. We would especially like to thank the following people for their support: Peter Fader, Elea Feit, and Charles Pensig at WCAI; Rob Christian, Janet Couperthwaite, and Lauren Fakonas at Russ Reid; and Tony DiPasquale, Sharron Silva, and John Wilburn at the American Red Cross.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*. 267–281.

Andreoni, J. 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal* **100**(401) 464–477.

Andreoni, J. 2006. Philanthropy. S.-C. Kolm, J. M. Ythier, eds., *Handbook on the Economics of Giving, Reciprocity and Altruism*, vol. 2. Elsevier, 1201–1269.

Ariely, D., A. Bracha, S. Meier. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *The American Economic Review* **99**(1) 544–555.

Arnett, D. B., S. D. German, S. D. Hunt. 2003. The identity salience model of relationship marketing success: The case of nonprofit marketing. *Journal of Marketing* **67**(2) 89–105.

Auten, G., D. Joulfaian. 1996. Charitable contributions and intergenerational transfers. *Journal of Public Economics* **59**(1) 55–68.

Bekkers, R. H. F. P., P. Wiepking. 2010. A literature review of empirical studies of philanthropy: eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly* **40**(5) 924–973.

Belloni, A., V. Chernozhukov. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**(2) 521–547.

Bennett, R. 2009. Impulsive donation decisions during online browsing of charity websites. *Journal of Consumer Behaviour* **8**(2-3) 116–134.

Bennett, R., R. Kottasz. 2000. Emergency fund-raising for disaster relief. *Disaster Prevention and Management* **9**(5) 352–360.

Bradić, J. 2014. Support recovery via weighted maximum-contrast subagging. *Arxiv preprint arXiv:1306.3494v3* URL `http://arxiv.org/pdf/1306.3494v3.pdf`.

Breeze, B., J. Dean. 2012. Pictures of me: user views on their representation in homelessness fundraising appeals. *International Journal of Nonprofit and Voluntary Sector Marketing* **17**(2) 132–143.

Breiman, L., P. Spector. 1992. Submodel selection and evaluation in regression. The x-random case. *International Statistical Review* **60**(3) 291–319.

Brown, P. H., J. H. Minty. 2008. Media coverage and charitable giving after the 2004 tsunami. *Southern Economic Journal* **75**(1) 9–25.

Brown, S., M. N. Harris, K. Taylor. 2012. Modelling charitable donations to an unexpected natural disaster: Evidence from the US Panel Study of Income Dynamics. *Journal of Economic Behavior & Organization* **84**(1) 97–110.

Bühlmann, Peter, Bin Yu. 2002. Analyzing bagging. *The Annals of Statistics* **30**(4) 927–961.

Bult, J. R., H. Van der Scheer, T. Wansbeek. 1997. Interaction between target and mailing characteristics in direct marketing, with an application to health care fund raising. *International Journal of Research in Marketing* **14**(4) 301–308.

Burnett, K. 2002. *Relationship Fundraising: A Donor-Based Approach to the Business of Raising Money (2nd ed.)*. John Wiley and Sons.

Calain, P. 2013. Ethics and images of suffering bodies in humanitarian medicine. *Social Science & Medicine* **98** 278–285.

Charng, H.-W., J. A. Piliavin, P. L. Callero. 1988. Role identity and reasoned action in the prediction of repeated behavior. *Social Psychology Quarterly* 303–317.

de Véricourt, F., M. S. Lobo. 2009. Resource and revenue management in nonprofit operations. *Operations Research* **57**(5) 1114–1128.

Desmet, P., F. M. Feinberg. 2003. Ask and ye shall receive: The effect of the appeals scale on consumers donation behavior. *Journal of Economic Psychology* **24**(3) 349–376.

Diamond, W. D., S. Gooding-Williams. 2002. Using advertising constructs and methods to understand direct mail fundraising appeals. *Nonprofit Management and Leadership* **12**(3) 225–242.

Doob, A. N., D. S. McLaughlin. 1989. Ask and you shall be given: Request size and donations to a good cause. *Journal of Applied Social Psychology* **19**(12) 1049–1056.

Efron, B. 2014. Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109**(507) 991–1007.

Falk, A. 2007. Gift exchange in the field. *Econometrica* **75**(5) 1501–1511.

Fan, J., F. Han, H. Liu. 2014. Challenges of big data analysis. *National Science Review* **1**(2) 293–314.

Fan, Y., R. Li. 2012. Variable selection in linear mixed effects models. *The Annals of Statistics* **40**(4) 2043–2068.

Fennis, B. M., L. Janssen, K. D. Vohs. 2009. Acts of benevolence: A limited-resource account of compliance with charitable requests. *Journal of Consumer Research* **35**(6) 906–924.

Fessler, J. A., A. O. Hero. 1994. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing* **42**(10) 2664–2677.

Frey, B. S., F. Oberholzer-Gee. 1997. The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review* **87**(4) 746–755.

Gneezy, U., A. Rustichini. 2000. Pay enough or don't pay at all. *Quarterly Journal of Economics* **115**(3) 791–810.

Hastie, T., R. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning*. Springer series in Statistics, New York, NY.

Homrighausen, D., D. J. McDonald. 2013. The lasso, persistence, and cross-validation. *Proceedings of the 30th International Conference on Machine Learning*.

Hyde, M. K., K. M. White. 2009. To be a donor or not to be? Applying an extended theory of planned behavior to predict posthumous organ donation intentions. *Journal of Applied Social Psychology* **39**(4) 880–900.

Jones, A., J. Posnett. 1991. Charitable donations by UK households: evidence from the Family Expenditure Survey. *Applied Economics* **23**(2) 343–351.

Karl, A. T., Y. Yang, S. L. Lohr. 2014. Computation of maximum likelihood estimates for multiresponse generalized linear mixed models with non-nested, correlated random effects. *Computational Statistics & Data Analysis* **73** 146–162.

Karlan, D., J. A. List. 2007. Does price matter in charitable giving? Evidence from a natural field experiment. *American Economic Review* **97**(5) 1774–1793.

Karlan, D., J. A. List, E. Shafir. 2011. Small matches and charitable giving: evidence from a natural field experiment. *Journal of Public Economics* **95** 344–350.

Kitchen, H. 1992. Determinants of charitable donations in Canada: a comparison over time. *Applied Economics* **24**(7) 709–713.

Klein, K. 2007. *Fundraising for social change*. John Wiley and Sons.

Kleiner, A., A. Talwalkar, P. Sarkar, M. I. Jordan. 2012. The big data bootstrap. *Proceedings of the 29th International Conference on Machine Learning*.

Kleiner, A., A. Talwalkar, P. Sarkar, M. I. Jordan. 2014. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society* **B76**(4) 795–816.

Lacetera, N., M. Macis, R. Slonim. 2012. Will there be blood? Incentives and displacement effects in pro-social behavior. *American Economic Journal: Economic Policy* **4**(1) 186–223.

Lafferty, B. A., R. E. Goldsmith, G.T.M. Hult. 2004. The impact of the alliance on the partners: A look at cause–brand alliances. *Psychology & Marketing* **21**(7) 509–531.

Laird, N. M., J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* **38**(4) 963–974.

Landry, C., A. Lange, J. A. List, M. K. Price, N. G. Rupp. 2006. Toward an understanding of the economics of charity: Evidence from a field experiment. *Quarterly Journal of Economics* **121**(2) 747–782.

Landry, C. E., A. Lange, J. A. List, M. K. Price, N. G. Rupp. 2010. Is a donor in hand better than two in the bush? Evidence from a natural field experiment. *American Economic Review* **100**(3) 958–983.

Lankford, R. H., J. H. Wyckoff. 1991. Modeling charitable giving using a Box-Cox standard Tobit model. *The Review of Economics and Statistics* **73**(3) 460–470.

Leszczyc, P.T.L., M.H. Rothkopf. 2010. Charitable motives and bidding in charity auctions. *Management Science* **56**(3) 399–413.

Lien, R. W., S. M. R. Iravani, K. R. Smilowitz. 2014. Sequential resource allocation for nonprofit operations. *Operations Research* **62**(2) 301–317.

List, J. A. 2011. The market for charitable giving. *Journal of Economic Perspectives* **25**(2) 157–180.

List, J. A., Y. Peysakhovich. 2011. Charitable donations are more responsive to stock market booms than busts. *Economics Letters* **110** 166–169.

McLachlan, G. J., T. Krishnan. 2008. *The EM Algorithm and Extensions (2nd ed.).* Wiley-Interscience.

Meer, J., H. S. Rosen. 2011. The ABCs of charitable solicitation. *Journal of Public Economics* **95**(5) 363–371.

Meier, L., S. Van De Geer, P. Bühlmann. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society* **B70**(1) 53–71.

Meinshausen, N., P. Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society* **B72**(4) 417–473.

Pedraza-Martinez, A. J., L. N. Van Wassenhove. 2013. Vehicle replacement in the International Committee of the Red Cross. *Production and Operations Management* **22**(2) 365–376.

Privett, N., F. Erhun. 2011. Efficient funding: Auditing in the nonprofit sector. *Manufacturing & Service Operations Management* **13**(4) 471–488.

PSID. 2012. U.S. Panel Study of Income Dynamics. `http://psidonline.isr.umich.edu/`.

Rudin, C., D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. N. Gross, B. Huang, S. Ierome, D. F. Isaac, A. Kressner, R. J. Pasonneau, A. Radeva, L. Wu. 2012. Machine learning for the New York City power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(2) 328–345.

Sargeant, A., E. Jay. 2004. *Building donor loyalty: the fundraiser's guide to increasing lifetime value*. John Wiley and Sons.

Sargeant, A., E. Jay, S. Lee. 2006. Benchmarking charity performance: returns from direct marketing in fundraising. *Journal of Nonprofit & Public Sector Marketing* **16**(1-2) 77–94.

Sargeant, A., J. Kähler. 1999. Returns on fundraising expenditures in the voluntary sector. *Nonprofit Management and Leadership* **10**(1) 5–19.

Sargeant, A., D. C. West, E. Jay. 2007. The relational determinants of nonprofit web site fundraising effectiveness. *Nonprofit Management and Leadership* **18**(2) 141–156.

Schaefer, R. L. 1986. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation* **25**(1-2) 75–91.

Schelldorfer, J., P. Bühlmann, S. van de Geer. 2011. Estimation for high-dimensional linear mixed-effects models using $\ell 1$-penalization. *Scandinavian Journal of Statistics* **38**(2) 197–214.

Schokkaert, E. 2006. The empirical analysis of transfer motives. S.-C. Kolm, J. M. Ythier, eds., *Handbook on the Economics of Giving, Reciprocity and Altruism*, vol. 1. Elsevier, 127–181.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**(2) 461–464.

Shang, J., R. Croson. 2009. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal* **119**(540) 1422–1439.

Shang, J., A. Reed, R. Croson. 2008. Identity congruency effects on donations. *Journal of Marketing Research* **45**(3) 351–361.

Shaw, D., E. Shiu, I. Clarke. 2000. The contribution of ethical obligation and self-identity to the theory of planned behaviour: An exploration of ethical consumers. *Journal of Marketing Management* **16**(8) 879–894.

Simmons, R. G., M. Schimmel, V. A. Butterworth. 1993. The self-image of unrelated bone marrow donors. *Journal of Health and Social Behavior* 285–301.

Small, D. A., G. Loewenstein. 2003. Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty* **26**(1) 5–16.

Small, D. A., G. Loewenstein, P. Slovic. 2007. Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes* **102**(2) 143–153.

Smithson, M., E. C. Merkle. 2014. *Generalized linear models for categorical and continuous limited dependent variables*. Chapman & Hall/CRC.

Terry, D. J., M. A. Hogg, K. M. White. 1999. The theory of planned behaviour: self-identity, social identity and group norms. *British Journal of Social Psychology* **38**(3) 225–244.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B58**(1) 267–288.

Tomasini, R. M., L. N. Van Wassenhove. 2009. From preparedness to partnerships: case study research on humanitarian logistics. *International Transactions in Operational Research* **16** 549–559.

Van Diepen, M., B. Donkers, P. H. Franses. 2009. Does irritation induced by charitable direct mailings reduce donations? *International Journal of Research in Marketing* **26**(3) 180–188.

Wang, H., G. Li, C.-L. Tsai. 2007. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B69**(1) 63–78.

Warwick, M. 2008. *How to write successful fundraising letters (2nd ed.)*. John Wiley and Sons.

WCAI. 2012. Cultivating disaster donors: A WCAI Research Opportunity sponsored by Russ Reid and the American Red Cross. `http://www.wharton.upenn.edu/wcai/files/Russ_Reid-ARC_Webinar.pdf`.

Wright, S. J., R. D. Nowak, M. A. T. Figueiredo. 2009. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing* **57**(7) 2479–2493.

Yen, S. T. 2002. An econometric analysis of household donations in the USA. *Applied Economics Letters* **9**(13) 837–841.

Zhang, Y., R. Li, C.-L. Tsai. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**(489) 312–323.

Zou, H., T. Hastie, R. Tibshirani. 2007. On the “degrees of freedom” of the lasso. *The Annals of Statistics* **35**(5) 2173–2192.