The Role of Time, Weather and Google Trends in Understanding and Predicting Web Survey Response

Qixiang Fang[1], Joep Burger[2], Ralph Meijers[2], and Kees van Berkel[2]

[1] *Utrecht University*
[2] *Statistics Netherlands*

Abstract

Significant efforts have been made to understand the effects of time-invariant factors (e.g. gender, personality and education) on web survey response. Time-invariant factors alone, however, cannot account for most variations in observed response patterns, especially fluctuations of response rates over time. This observation inspires us to investigate the counterpart of time-invariant factors, namely time-varying factors and the potential roles they play in web survey response processes. Specifically, we study the effects of time, weather and societal trends (derived from Google Trends data) on the daily responses of the web mode of the 2016 and 2017 Dutch Health Surveys. We find, among others, that weekends, holidays, pleasant weather, disease outbreaks and terrorism salience are associated with fewer responses. Furthermore, our privacy-friendly modelling approach achieves satisfactory prediction accuracy of both daily and cumulative response rates.

Keywords: web survey; response rates; weather; Google Trends; survival analysis

1 Introduction

Despite their popularity, web surveys suffer from lower response rates and consequently, compromised data quality (i.e. sometimes higher bias and always lower precision) (Taylor & Scott, 2019). To understand the underlying mechanisms of web survey response, significant research efforts have been made. To date, research has primarily focused on the effects of "time-invariant" factors, whose values stay relatively (or presumably) constant during a survey's data collection period for many potential respondents. Examples are personal traits and background (e.g. age, personality, education), regional characteristics (e.g. degree of urbanisation, population density) and survey design features (e.g. survey length, question wording).

Such findings have greatly improved our understanding about web survey response processes and inspired survey design efforts aiming at increasing response rates. However, time-invariant factors alone can explain only a limited proportion of variations in response patterns (e.g. Erdman and Bates, 2017). This can be because with constant values, these factors cannot properly account for temporal fluctuations of survey response rates (especially within the same or between similar surveys targeting similar populations), for instance, across different days of a week (Faught et al., 2004), months (Losch et al., 2002) and years (Sheehan, 2006). This observation directs us to look into the counterpart of time-invariant factors,

namely time-varying factors. Specifically, we investigate whether and how various time-varying factors (including time, weather and societal trends) may explain and predict the daily response patterns of the web mode of the 2016 and 2017 Dutch Health Surveys.

2 Time-Varying Contextual Factors

2.1 Definitions

Time-varying factors are variables whose values may differ over time (Singer & Willett, 2003). Some change values naturally, others by design. Examples of potentially influencing time-varying factors of survey response are: personal availability, individual emotional status, day of a week, weather, holidays, and societal trends (e.g. public concerns about data privacy).

In this paper, we focus on time-varying factors that are also contextual factors. Contextual factors are variables that cannot be influenced by study participants because they are largely or totally determined by external stochastic processes (Kalbfleisch & Prentice, 2002). They typically assess changing characteristics of the physical or social environment where study participants live. Factors like weather, time and societal trends fit this definition.

We focus on contextual time-varying factors for two practical considerations. First, data for contextual factors are usually publicly available, non-personal and thus circumvent privacy issues, which is a noteworthy advantage given that using personal data for research purposes has become more difficult because of growing data privacy concerns among the public and the introduction of more stringent privacy regulations. Second, because contextual factors are (largely) free from the influence of study subjects, there is little concern for reverse causality and hence more internal validity for the study.

2.2 Literature Review

There are two types of survey response studies on time-varying contextual factors. The first focuses on the effect of time, such as years, seasons, months and days of a week. For instance, Sheehan (2006) analysed 31 web surveys and concluded that the *year* in which a survey was published was the most crucial predictor of response rates. Losch et al. (2002) found that completing a survey interview during *summer* in Iowa (US) required more contact attempts than in other seasons. Similarly, Göritz (2014) documented that members of a German online panel were more likely to start and finish studies in *winter* than any other season. Contrasting these two findings, Svensson et al. (2012) in a Swedish sample found that the highest response rate was in *September*. Faught et al. (2004) noted in their experiment on US manufacturers that survey response rates were the highest when the email invitation was sent on either *Wednesday morning* or *Tuesday Afternoon*. Contrary to this, Sauermann and Roach (2013) conducted experiments among US researchers and did not find the timing of the e-mail invitation (in terms of the day of a week) to result in different response rates in a web survey. However, they did find that people responded less in the *weekend* and would postpone their responses until the next week.

The second type of studies concerns the influence of weather. For example, Potoski et al. (2015) analysed eight surveys from 2001 to 2007 and showed that on unusually *cold* and *warm* days, wealthier people are more likely to participate in surveys. Cunningham (1979) found more *pleasant weather* (i.e. more sunshine, higher temperature, lower humidity) to improve a person's willingness to participate in an interview. The effect of *weather* on survey participation, however, likely goes beyond these two findings. Simonsohn (2010) showed that on *cloudier* days people engage more in academic activities, which share some common characteristics with survey participation (e.g. high cogni-

tive load and low immediate returns). Keller et al. (2005) showed that higher air pressure has a positive influence on mood, which in turn may lead to increased helping behaviour (e.g. fulfilling a survey request) (Weyant, 1978). Therefore, cloudiness and air pressure may also impact survey response decisions.

2.3 Proposed Factors

These previous studies suggest that time-varying contextual factors likely affect survey response. However, their findings likely only apply to specific settings (e.g. country, survey mode) and some contradict one another. Therefore, we recognise a need for more research and we propose the following time-varying contextual factors for this study (see Appendix A for an overview).

2.3.1 Time

The first two time-varying contextual factors relate to time, including *the day of a week* and *public holidays*. The former is chosen because its effect on survey responses is still unclear. With the latter factor, we hypothesise that during holidays, people travel around more, spend more time with family or want to rest and are consequently less responsive to survey requests.

2.3.2 Weather

We use 20 variables that measure daily temperature, sunshine, precipitation, wind, cloud, visibility, humidity and air pressure in various ways (e.g. maximum, minimum and average).

2.3.3 Societal Trends

We hypothesise that societal trends including *disease outbreaks*, *public privacy concerns*, *public outdoor engagement* and *terrorism salience* may influence survey participation and elaborate on each below.

Disease Outbreaks Medical conditions such as the common cold, the flu, hay fever and depression tend to affect a large number of individuals, especially during certain times of a year. When sick, the individuals may stay at home more and reduce outdoor and professional activities. They may lack the resources to engage in cognitively demanding activities and helping behaviour. These physical, behavioural and psychological changes may in turn impact survey participation. Therefore, we hypothesise that disease outbreaks are a potential influencing factor of survey response.

Privacy Concerns Two studies on the 1990 and 2000 U.S. census found that an increase in concern about privacy and confidentiality issues was consistently associated with less census participation (Singer et al., 1993; Singer et al., 2003). Two more recent studies also linked greater privacy concerns to more unit and item non-response (Dahlhamer et al., 2008; Bates et al., 2008). Given these findings, we hypothesise that the level of general societal concerns about data privacy are negatively associated with survey response rates.

Public Outdoor Engagement When people are engaged in outdoor activities (e.g. public events, holidays, travel), it is conceivable that they are less likely to participate in surveys because, for instance, they are more difficult to reach or the surveys are not smartphone-friendly (like the surveys in this study). Therefore, we hypothesise that higher levels of outdoor engagements negatively affect concurrent survey response rates.

Terrorism Salience Terrorist events have impactful individual consequences. For instance, higher levels of terrorism salience are linked to more negative emotions (Fischer et al., 2006), worse mental health (Fischer & Ai, 2008), more media consumption (Lachlan et al., 2009), increased contact with family and friends

(Goodwin et al., 2005), irrational travel behaviour (Baumert et al., 2019), lower social trust (Geys & Qari, 2017) and less occupational networking activities (Kastenmüller et al., 2011). Therefore, terrorism salience can indirectly influence survey response behaviour by inducing changes in, for example, health status, media use, travel behaviour and interpersonal relationships. We tentatively hypothesise an overall negative effect of terrorism salience on survey response rates. Furthermore, terrorism salience can especially relevant to the Dutch context, considering that the Netherlands and its nearby countries (e.g. Germany and France) have suffered from terrorist threats, attacks or related events and issued terrorism warnings during the recent years (see US Department of State, 2017; US Department of State, 2018).

3 Research Aims

The first goal of the current study is to investigate whether and how daily time-varying contextual factors including day of a week, holidays, weather and societal trends may influence daily web survey response. The second goal is to evaluate the predictive performances of the resulting models and predictor estimates, as we believe that models and predictors are generally more useful when they not only explain the training data but also generalise to unseen observations.

Note that we purposefully avoid using dummy variables of months and seasons in our models, partly because we expect the daily time-varying variables to capture most monthly and seasonal trends and partly because there is not enough data variation to ensure reliable estimation of the relevant month and season effects.

4 Data

4.1 The Dutch Health Surveys

We analyse the response decisions of individuals who were invited to the web mode of the 2016 or the 2017 Dutch Health Survey, a yearly survey aiming to assess the developments in health, medical contacts, lifestyle and preventive behaviour of the Dutch population. The sampling frame comprises of persons of all ages residing in private households.

A yearly sample is divided into 12 cohorts, each corresponding to a data collection period (DCP) that lasts about a month. A DCP begins with the delivery of a physical invitation letter to the sample unit's home with a request to respond to the online survey using a computer or tablet (no smartphone). In case of no response from the person after about one week, up to two physical reminder letters would follow (at an interval of one week). Both invitation and reminder letters contain a web link to the survey and a unique login code.

The 2016 and 2017 Dutch Health Surveys share a similar design, making comparison and integration of data from both years valid and simple. Table 1 presents the expected delivery days of the letters, sample sizes and response rates in both years. Note that the expected delivery days of the letters differ within and between the years.

Table 1: Comparison of the 2016 and 2017 Dutch Health Survey (Web Mode)

	2016	**2017**
Expected Delivery Day of Invitation	Fri. (Jan.-Jun.) Sat. (Jul.-Dec.)	Thur.
Expected Delivery Day of Reminder	Fri. (Jan.-Jun.) Sat. (Jul.-Dec.)	Sat.
Sample Size	15007	16972
Response Rate	34.8%	34.2%

4.2 Weather Data

The Royal Netherlands Meteorological Institute (KNMI) records average daily temperature, sunshine, precipitation, wind, cloud, visibility, hu-

midity and air pressure in the Netherlands. We retrieved the 2016 and 2017 records (20 variables in total, see Appendix A) from the KNMI website (KNMI, 2019).

4.3 Societal/Google Trends

The four societal trends of interest are *disease outbreaks*, *privacy concerns*, *public outdoor engagement* and *terrorism salience*. To our knowledge, there is no publicly available data on these societal trends on a daily scale in the Netherlands. Our solution is Google Trends (GT).

GT offers periodical summaries of user search data for many regions and for any possible search term. These summaries, available as indices, represent the number of Google searches that include a given search term in a specified period (e.g. day, week or month). The data are scaled for the requested period between 0 and 100, with 0 indicating no search at all and 100 the highest search volume in that period. These indices can offer (almost) real-time insights into various aspects of human societies, such as disease outbreaks (Carneiro & Mylonakis, 2009), economic performance (Choi & Varian, 2012) and salience of immigration issues and terrorism (Mellon, 2014).

To ensure the validity of the GT data, we followed the recommendation of Zhu et al. (2012) to filter out terms with low validity. Furthermore, we applied a resampling and calibration procedure to improve the reliability of and comparability of the GT data (see Appendix B for more details).

The resulting search terms (with good validity and reliability) are as follows. For *disease outbreaks*, we use the Dutch terms for "flu", "cold" and "depression". For *privacy concerns*, we use the Dutch terms for "data leaks" and "hacking". For *outdoor engagement*, we used two Dutch terms indicating traffic jams and interests in festivals. Lastly, for *terrorism salience*, we used the term "terrorist". Note that we hypothesised these search terms prior to any analysis, thus reducing the risk of spurious correlations.

5 Methods

5.1 Discrete-Time Survival Analysis

Our research questions and data call for discrete-time survival analysis (Singer & Willett, 2008) which can 1) model the transition from non-response to response, 2) incorporate time-varying predictors, and 3) work with discrete-time data (on a daily scale).

Discrete-time survival analysis uses *hazard* to assess the risk of event occurrence, which in this research refers to the *conditional probability* of a person responding to the survey on a given day, given the individual did not respond earlier. Since hazards are probabilities restricted to the interval [0, 1], we can use a binomial link function to connect the linear predictors and the outcome hazards. Under this binomial regression model, the exponential term of a parameter estimate quantifies the change in the conditional odds per unit difference in the predictor.

Discrete-time survival models require specifying the baseline hazard for every time point. In this study, an individual become "at risk" (of responding to the survey) after receiving the invitation letter. Thus, we conceptualise each time point as a linear combination of the number of days since the delivery of the previous invitation or reminder letter (i.e. *days*, continuous) and the specific survey phase this time point is in (*survey phase*, categorical with three levels: "invitation", "reminder1" and "reminder2").

Note that we define the "*days*" variable as the number of days since the last letter instead of the invitation letter, because this removes dependency between the *days* and *survey phase* variables. For instance, "Day 5" in combination with "survey phase: reminder1" indicates that this is day 5 since the first reminder arrived.

5.2 Lasso Regularisation

When used with a large number of highly correlated (weather and GT) predictors, binomial regression would result in a complex, overfit model where it is difficult to interpret parameter estimates and determine relevant predictors.

Our solution is Lasso (regularisation), a machine learning technique capable of performing variable selection (while achieving good prediction). Generally speaking, Lasso works by adding a penalty term λ to the likelihood function of a model, which shrinks parameter estimates towards zero. In doing so, Lasso retains only a small number of relevant variables (i.e. the ones with non-zero estimates) and thus produces a more parsimonious and interpretable model. Because this variable selection procedure is automatic, we can also avoid using the increasingly controversial p-values and confidence intervals to judge the relevance of a variable. In addition, by introducing a small bias to the model, Lasso significantly reduces model variances and thereby improves the model's out-of-sample predictive performance.

The value of λ needs to be carefully selected, because up until a certain point, the increase in λ is beneficial as it only reduces the variance (and hence avoids overfitting), without losing any important properties in the data. After a certain threshold, however, the model starts losing important properties, giving rise to bias in the model and thus underfitting. To find the optimal λ, we used 10-fold cross-validation, as recommended by Hastie et al. (2009).

5.3 Data Partition

The data was split into a *training set* and a *test set* in such a way that ensures a good trade-off between sufficient variation in the training data and independence of the test data from the training data. The resulting training set comprises the complete 2016 Dutch Health Survey observations and the first half of the 2017 data, while the test set contains the rest (i.e. the last six DCPs in 2017).

5.4 Three Models and Evaluation

We fitted three models to the training data, evaluated and compared their predictive performances on the test data. These models are 1) "baseline model", which includes only the baseline hazard predictors ("days" and "survey phase"); 2) "full model", which additionally includes all the time-varying contextual predictors; 3) lastly, "interaction model".

With the full model, we assume that the effects of the predictors do not vary with time. Such a model is more parsimonious and easier to interpret. However, in reality, the effect of a predictor may depend on time. The "interaction model" allows for this possibility, where interaction terms between the baseline hazard predictors and the time-varying contextual predictors are included. Note that the resulting model contains non-zero interaction terms whose corresponding main effects are shrunk to zero. Because of this, we do not interpret the model's parameter estimates but only assess its predictive performance.

We use *root mean squared error* (RMSE) as the evaluation criterion, which quantifies the distance between the observed and the predicted daily hazards. Furthermore, we plotted the predicted cumulative response rates of the three models against the observed ones.

5.5 Variable Importance

We further assessed the importance of each predictor by calculating the increase in the model's prediction error (i.e. RMSE) after permuting the variable (Fisher et al., 2019). Permuting the variable breaks the relationship between the variable and the true outcome. Thus, a variable is "important" if shuffling its values increases the model's prediction error, because in this case the model relies on this variable for better prediction. A variable is "unimportant" if permuting its values leaves the model error unchanged

or worse. For accurate estimates, we used 20 permutations per variable and averaged the resulting variable importance scores.

5.6 Research Repository

All the data and codes are available at https://doi.org/10.5281/zenodo.4159915.

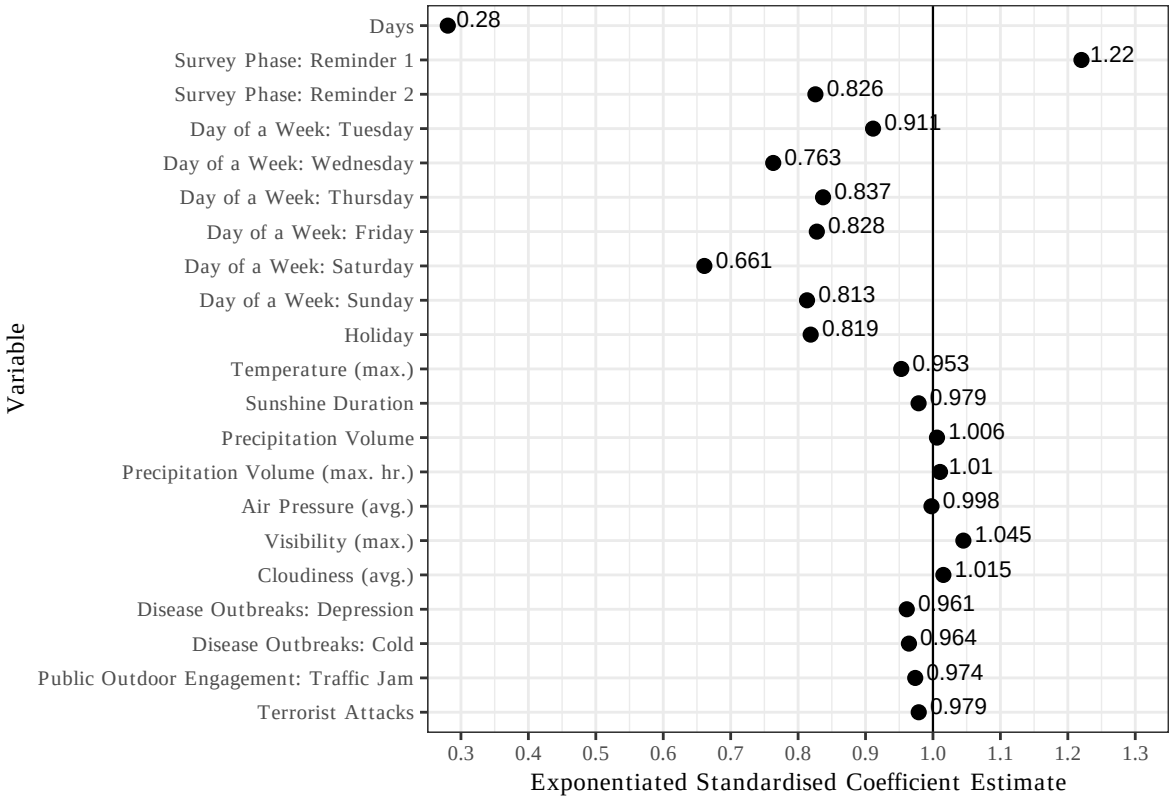

Figure 1: Exponentiated Standardised Estimates of Predictors in the Full Model

6 Results

6.1 Model Estimates and Interpretation

Figure 1 shows the exponentiated standardised parameter estimates of the predictors considered non-zero and thus relevant by the full model. The estimates quantify how much one SD change in the predictors affects the conditional odds of response on a given day, all else being constant.

Both variables that define the baseline hazards are strong predictors of survey response. Specifically, *days* has the largest effect on survey responses among all. An estimate of about 0.28 suggests that for one SD increase in the number of *days* (about 5.34 days) since the last letter, the conditional odds of a person responding to the survey on that day are reduced by about 72%. *Survey phase* is also a relevant predictor: the first reminder letter increases the conditional odds of response by about 22% than in the invitation phase, while the second reminder letter lowers the conditional odds by about 17%.

Turning to the time-varying contextual predictors: *Day of a week* appears as a strong predictor. Compared to Monday, all non-Mondays are associated with fewer responses. Saturday shows the strongest negative effect, which lowers the conditional odds of response by about 34%. The effect of Sunday on response odds is also non-negligible and negative.

Holiday also has a negative effect on re-

sponses, reducing the conditional response odds by about 18% compared to non-holidays.

Despite having smaller effect estimates than the previous variables, the weather variables show a clear pattern. When the weather is nicer (e.g. higher temperature, longer sunshine duration, less rain, higher air pressure, and less cloudy), the conditional response odds are also lower. For one SD change in these weather variables, the conditional response odds can change by a maximum of about 5%. The only exception to this pattern is the variable *maximum visibility*, which has a positive coefficient.

Similar to the weather variables, the GT variables have, if not zero, small effects. Among the disease outbreaks variables, "depression" and "cold" show negative effects on survey responses, while the other indicators ("flu", "hay fever" and "influenza") are not retained by the model. The two variables concerning data privacy concerns, namely "data leak" and "hacking", have also been left out by the model. Between the two variables concerning public outdoor engagement, "traffic jam" negatively predicts survey responses, while "festival" has no effect. Finally, "terrorist" also has a small negative influence on survey responses. Note that the interpretation of the GT variables in terms of their effect sizes is difficult and can be misleading, because the variables are measured on less well-defined scales.

6.2 Model Performance

The RMSE scores of the baseline, full and interaction models are 0.005528, 0.005274 and 0.004738, respectively. That is, the full model (compared to the baseline model) and the interaction model (compared with the full model) reduce prediction error by about 4.6% and 10%.

Figure 2 shows the predicted cumulative response rates of the three models across all survey phases in the test data, against the observed scores. In the invitation phase, the interaction model achieves the best prediction performance among the three, especially after day 4 where the prediction is almost perfect. Both the full model and the baseline model underpredict cumulative response rates.

In the first reminder phase, all models appear to predict cumulative response rates better than in the invitation phase. The interaction model again performs the best, followed by the full model and lastly, the baseline model.

The beginning and the end of the second reminder phase see all of the three models predict well. However, all three largely overpredict cumulative response rates in the mid-period, with the interaction model faring only slightly better than the others.

6.3 Variable Importance

Figure 3 shows the mean variable importance scores of the predictors retained by the full model, sorted in descending order. A score above 1 indicates positive contribution to prediction accuracy, while a score below 1 suggests the opposite. *Days* contributes the most to prediction accuracy, followed by the two weekend indicators (i.e. *Saturday* and *Sunday*) and subsequently, the two *survey phase* indicators. The weekday indicators, except for Thursday, all contribute to prediction accuracy. Most *weather* variables contribute to prediction accuracy, despite some scoring only slightly above 1. The *holiday* variable has a very small positive contribution. All of the GT variables, *maximum visibility* and the *Thursday* indicator have negative variable importance scores.

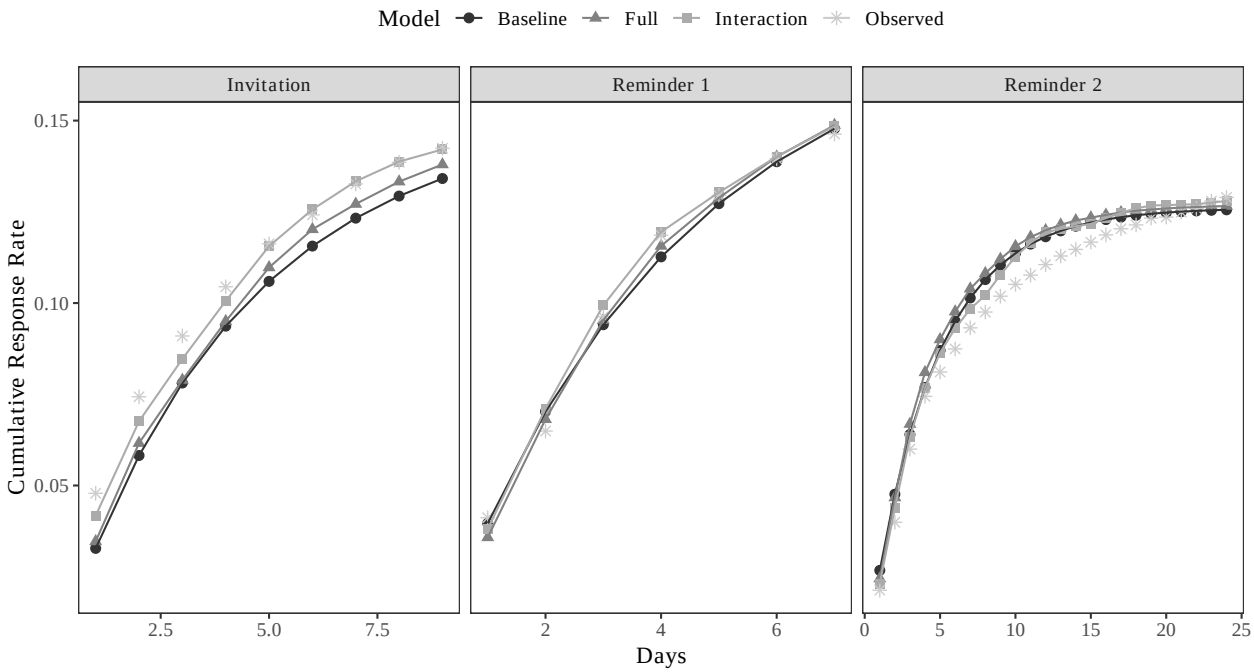

Figure 2: Predicted and Observed Cumulative Response Rate by Survey Phase and Model

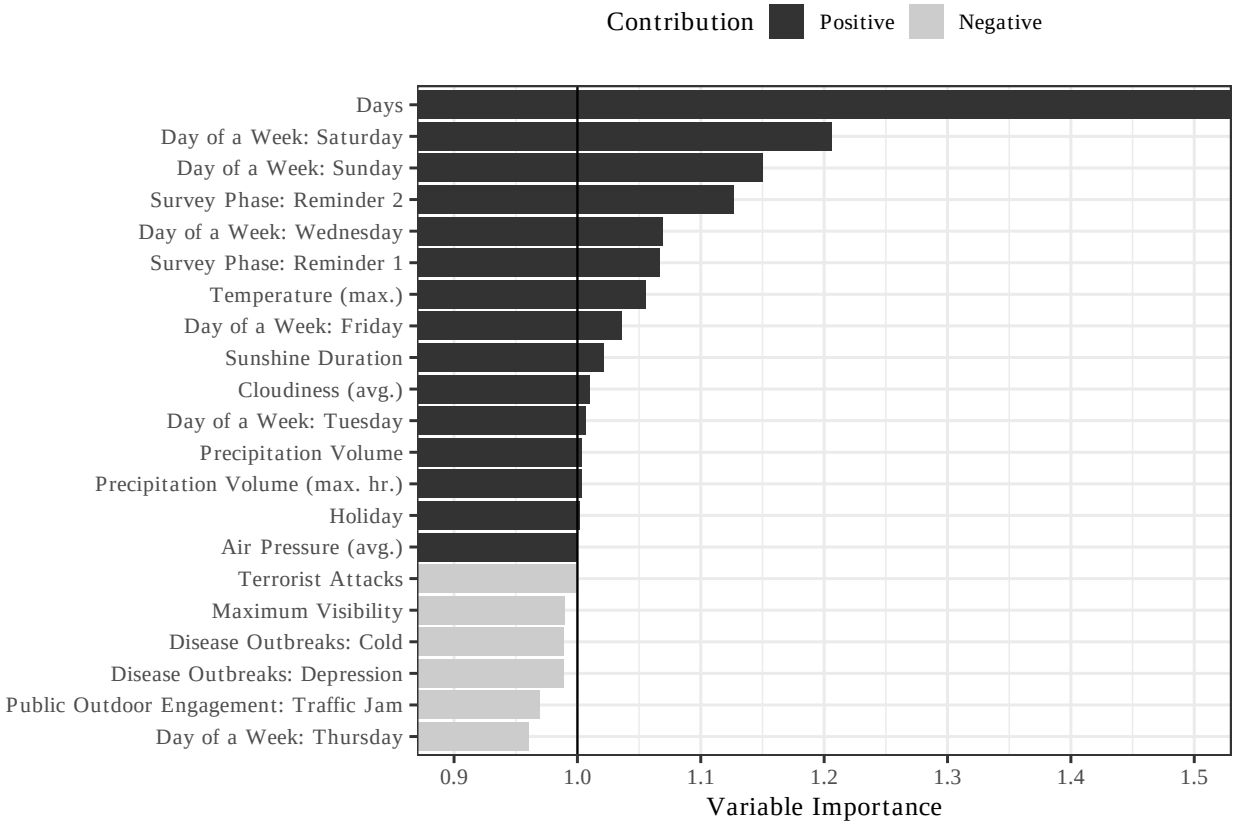

Figure 3: Variable Importance of Predictors in the Full Model

7 Discussion & Conclusion

7.1 Effects of Time-Varying Contextual Predictors

We see that *Monday* is the day of a week most positively associated with survey response. All the other days have moderate to strong negative effects on response compared to Monday. Among them, Saturday and Sunday show the strongest and the third largest negative effects, respectively. This result is consistent with the previous finding by Sauermann and Roach (2013) that people would postpone their responses to surveys administered in the weekend until the next week.

Furthermore, a potential reason for Saturday's strongest negative effect might be because on Saturday, people have the greatest need among the week to rest and have more family and household obligations. The work of Roeters (2018) about time use in 2016 in the Netherlands supports this speculation. The report shows that Saturday was the day among the week when Dutch people spent the most time on household and family care and leisure activities.

Holidays also negatively predicts response, which agrees with the finding of Losch et al. (2002) that more contact attempts are required in the holiday seasons. The underlying reason might be similar to the negative effect of Saturday: People spend more time on leisure activities (e.g. travel and personal care) and family obligations and hence less on surveys.

Despite somewhat smaller effects, the estimates of the weather variables suggest that with more pleasant weather (e.g. higher temperature, longer sunshine duration, less rain, higher air pressure and less cloudy), response rates are lower. This is perhaps because people are more likely outdoors when the weather is nice and therefore unable to fill in a smartphone-unfriendly survey. This result contrasts the previous finding by Cunningham (1979) that more pleasant weather makes people more likely to accept a face-to-face interview request. Speculatively, the influence of weather on survey response might be moderated by survey modes.

Among the societal trends variables, both "depression" and "cold" negatively predict responses, which agrees with our hypothesis that when people are sick, they are less likely to respond. "Traffic jam", one of the two variables intended to measure public outdoor engagement, is negatively associated with response. This may suggest that on days when more people travel (by car), response rates are lower. The "festival" variable was not retained by the model, which may be because this search term is less indicative public outdoor engagement than is "traffic jam". Both variables ("data leak" and "hacking") measuring public concerns about data privacy have zero coefficients. This may suggest that in the Netherlands, participation in official surveys like the Dutch Health Survey is hardly influenced by data privacy concerns. Terrorism salience is negatively linked to response, supporting our hypothesis and previous findings about the negative individual and societal consequences of terrorism.

7.2 Model Predictive Performance

All three fitted models achieve acceptable predictive performances, indicated by their RMSE scores and estimated cumulative response rates. A reduction in RMSE up to 14.2% (between the baseline and the interaction model) is an incredible outcome. This would allow researchers to estimate response rates at the stage of survey planning much more accurately, which can mean a significant amount of resources saved.

All of the models moderately overpredict cumulative response rates in the mid-period of the second reminder phase. This requires further examination into, for instance, additional influencing factors of web survey responses and possible nonlinear relationships, for better response prediction.

Among the three models, the interaction model consistently performs the best, followed

by the full model and lastly, the baseline model. This speaks of the relevance of including time-varying contextual predictors like *day of a week*, *holiday* and *weather* for the understanding and the prediction of survey response. If prediction is the main goal, the interaction model (or even more complex "black-box" models) should be preferred (at the cost of interpretability). In contrast, if the goal is to explain, then the more interpretable and parsimonious full model should be preferred.

7.3 Variable Predictive Performance

Days is the single most important (i.e. predictive) predictor of survey response. *Survey phase*, *holidays* and most *day of a week* and *weather* variables also contribute to prediction accuracy with varying degrees. Two exceptions are the *maximum visibility* and the *Thursday* predictors, which lead to more prediction error. The first one suggests that *maximum visibility* might be an irrelevant predictor of survey response. Considering also its positive coefficient which disagrees with the overall negative effects of good weather on survey response, *maximum visibility* should preferably be left out. The negative variable importance of *Thursday* indicates that Thursday's effect estimate lacks generalisability. However, the *day of a week* variable as a whole still contributes strongly and positively to model prediction accuracy.

Some predictors have variable importance scores only slightly above 1, such as *Tuesday*, *precipitation volume (max. hr.)*, *holiday* and *air pressure (avg.)*. There can be different reasons for this. For instance, it might be because we do not have enough data for the model to obtain for these particular variables estimates that generalise better to unseen data. Or these variables might be, indeed, less relevant for survey response prediction. Drawing a convincing conclusion would require future research.

Lastly, all of the GT variables lead to worse prediction, likely due to the fact that GT data are scaled somewhat arbitrarily or too noisy. However, just because they do not contribute positively to prediction error in this study, it does not necessarily follow that the use of GT variables is always unwarranted or that the found effects of the GT variables are untrustworthy. Instead, GT data may be simply more suitable for more exploratory research, unless a better method for deriving measures from GT data becomes available.

7.4 Implication for Data Collection

Our results suggest that response rates are likely lower in the following circumstances: during the weekend (especially Saturday), summer (when the weather is generally nice and many people take holidays), public holidays and disease outbreaks (e.g. cold and depression).

Furthermore, response odds are highly negatively associated with the number of days into a survey phase, suggesting that securing as many responses as possible already in the beginning of a survey phase might be crucial for achieving a high final response rate. It might be also beneficial during survey planning to factor in the potential effects of relevant time-varying contextual factors on (the initial) survey responses.

Nevertheless, we need to be careful with causal claims. Despite reverse relationships being impossible here, this study is observational in nature and therefore, we can only tentatively suggest that survey researchers might want to implement their survey projects in times when response rates tend to be higher. More confidence in such advice can only be warranted by experiments.

Lastly, our findings and recommendations may only apply to web surveys whose design, target population and survey culture are similar to the web mode of the 2016 and 2017 Dutch Health Surveys. For others, we recommend conducting response modelling research based on past data of the same or similar surveys, using our analytic approach, which should be feasible

considering no involvement of personal or highly inaccessible data.

7.5 Implication for Survey Research

Our study uses discrete-time survival analysis combined with Lasso regularisation, which requires no personal data, results in interpretable models and achieves satisfactory prediction performance of response rates using only time-varying contextual predictors. Therefore, we highly recommend this privacy-friendly, interpretable and predictive modelling framework to survey researchers. Note that discrete-time survival analysis can also incorporate both time-invariant and time-varying predictors at the same time. See Singer and Willett, 2008 for more information.

References

Bates, N., Dahlhamer, J., & Singer, E. (2008). Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. *Journal of Official Statistics*, *24*(4), 591–612.

Baumert, T., de Obesso, M. M., & Valbuena, E. (2019). How does the terrorist experience alter consumer behaviour? An analysis of the Spanish case. *Journal of Business Research*, *115*, 357–364.

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557–1564.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*(1), 2–9.

Cunningham, M. R. (1979). Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan. *Journal of Personality and Social Psychology*, *37*(11), 1947–1956.

Dahlhamer, J. M., Simile, C. M., & Taylor, B. (2008). Do you really mean what you say? Doorstep concerns and data quality in the National Health Interview Survey (NHIS), In *Section on survey research methods – jsm 2008*.

Erdman, C., & Bates, N. (2017). The Low Response Score (LRS). *Public Opinion Quarterly*, *81*(1), 144–156.

Faught, K., Whitten, D., & Green, K. (2004). Doing survey research on the internet: Yes, timing does matter. *Journal of Computer Information Systems*, *44*(3), 26–34.

Fischer, P., & Ai, A. L. (2008). International terrorism and mental health: Recent research and future directions. *Journal of Interpersonal Violence*, *23*(3), 339–361.

Fischer, P., Greitemeyer, T., Kastenmüller, A., Jonas, E., & Frey, D. (2006). Coping with terrorism: The impact of increased salience of terrorism on mood and self-efficacy of intrinsically religious and nonreligious people. *Personality and Social Psychology Bulletin*, *32*(3), 365–377.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(117), 1–81.

Geys, B., & Qari, S. (2017). Will you still trust me tomorrow? The causal effect of terrorism on social trust. *Public Choice*, *173*(3-4), 289–305.

Goodwin, R., Willson, M., & Stanley, G. (2005). Terror threat perception and its consequences in contemporary Britain. *British Journal of Psychology*, *96*(4), 389–406.

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies (M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas, Eds.). In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY, Springer New York.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ, USA, John Wiley & Sons, Inc.

Kastenmüller, A., Greitemeyer, T., Aydin, N., Tattersall, A. J., Peus, C., Bussmann, P., Fischer, J., Frey, D., & Fischer, P. (2011). Terrorism threat and networking: Evidence that terrorism salience decreases occupational networking. *Journal of Organizational Behavior*, *32*(7), 961–977.

Keller, M. C., Fredrickson, B. L., Ybarra, O., Côté, S., Johnson, K., Mikels, J., Conway, A., & Wager, T. (2005). A warm heart and a clear head: The contingent effects of weather on mood and cognition. *Psychological Science*, *16*(9), 724–731.

KNMI. (2019). Daggegevens van het weer in Nederland [Daily weather data in the Netherlands]. http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi

Lachlan, K. A., Spence, P. R., & Seeger, M. (2009). Terrorist attacks and uncertainty reduction: Media use after September 11. *Behavioral Sciences of Terrorism and Political Aggression*, *1*(2), 101–110.

Losch, M. M. E. M., Maitland, A., Lutz, G., Mariolis, P., Gleason, S. C., & Aaron Maitland. (2002). The effect of time of year of data collection on sample efficiency: An analysis of behavioral risk factor surveillance survey data. *Public Opinion Quarterly*, *66*(4), 594–607.

Mellon, J. (2014). Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties*, *24*(1), 45–72.

Potoski, M., Urbatsch, R., & Yu, C. (2015). Temperature biases in public opinion surveys. *Weather, Climate, and Society*, *7*(2), 192–196.

Roeters, A. (2018). *Time use in the Netherlands* (tech. rep.). The Netherlands Institute for Social Research (SCP). The Hague.

Sauermann, H., & Roach, M. (2013). Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, *42*(1), 273–286.

Sheehan, K. B. (2006). E-mail survey response rates: A review. *Journal of Computer-Mediated Communication*, *6*(2), JCMC621.

Simonsohn, U. (2010). Weather to go to college. *Economic Journal*, *120*(543), 270–280.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. census. *Public Opinion Quarterly*, *57*(4), 465–482.

Singer, E., Van Hoewyk, J., & Neugebauer, R. J. (2003). Attitudes and behavior the impact of privacy and confidentiality concerns on participation in the 2000 census. *Public Opinion Quarterly*, *67*(3), 368–384.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, Oxford University Press.

Singer, J. D., & Willett, J. B. (2008). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, *18*(2), 155–195.

Svensson, M., Svensson, T., Hansen, A. W., & Lagerros, Y. T. (2012). The effect of reminders in a web-based intervention study. *European Journal of Epidemiology*, *27*(5), 333–340.

Taylor, T., & Scott, A. (2019). Do physicians prefer to complete online or mail surveys? Findings from a national longitudinal survey. *Evaluation and the Health Professions*, *42*(1), 41–70.

US Department of State. (2017). *Country Reports on Terrorism 2016 - The Netherlands* (tech. rep.). https://nl.usembassy.gov/2016-country-reports-terrorism-netherlands/

US Department of State. (2018). *Country Reports on Terrorism 2017 - The Netherlands* (tech. rep.). https://www.refworld.org/docid/5bcf1f7a13.html

Weyant, J. M. (1978). Effects of mood states, costs, and benefits on helping. *Journal of Personality and Social Psychology*, *36*(10), 1169–1176.

Zhu, J. J. H., Wu, L., Wang, X., & Qin, J. (2012). Assessing public opinion trends based on user search queries: Validity, reliability, and practicality. *World Association for Public Opinion Research*, 1–7.

Appendix A

Table: Overview of Time-Varying Variables

Variable	**Description**
Days	The number of days since the last invitation or reminder letter.
Survey Phase	Three categories: "invitation" (reference), "reminder 1", "reminder 2".
Day of a Week	Seven categories: from "Monday" (reference) to "Sunday".
Holiday	A dummy {0,1} indicating Dutch holidays, including New Year's Day, Carnival, Good Friday, Easter, Easter Monday, King's Day, Liberation Day, Ascension Day, Pentecost, Whit Monday, Saint Nicholas's Day, Christmas, and New Year's Eve.
Temperature	Three variables: daily maximum, minimum and average temperature (in °C).
Sunshine Duration	Daily sunshine duration (in hours)
Sunshine Percentage	Daily percentage of sunshine
Precipitation Volume	Two variables: daily total precipitation volume and hourly maximum precipitation volume (in mm)
Precipitation Duration	Daily precipitation duration (in hours)
Wind Speed	Three variables: daily maximum, minimum and average hourly wind speed (in m/s).
Cloudiness	Ordinal scale from 1 (low) to 9 (high), indicating the daily degree of average cloudiness.
Visibility	Two variables: daily minimum and maximum distance of visibility (in m).
Humidity	Three variables: daily maximum, minimum and average humidity (in %).
Air Pressure	Three variables: daily maximum, minimum and average air pressure (in hPa).
Disease Outbreaks	Five search terms (English in parentheses): "depressie" (depression), "griep (flu)", "influenza (influenza)", "verkoudheid (cold)" and "hooikoorts (hay fever)".
Data Privacy Concerns	"datalek (data leak)" and "hacking (hacking)".
Outdoor Engagement	"files (traffic jam)", "festival (festival)".
Terrorism Salience	"terrorist (terrorist)".

Appendix B

We used the following resampling and calibration procedure to ensure the reliability of the obtained GT data, taking into account the data retrieval limits imposed by the GT server (i.e. at most 244 daily GT scores per inquiry per search term). We omit the reliability test results of the raw and the calibrated GT scores due to word count limits.

We denote one GT query as $GT_s := \{G_{s,i}, i \in \{1:244\}\}$ for sample s, where G_{s1} is the index score of the first day in the requested period and G_{s244} the last day. The index s runs from 1 to 974 with $s = 1$ for the sample that starts on $t_1 :=$ 2015-05-03 and $s = 974$ for the sample that starts on $t_{974} :=$ 2017-12-31. The first day of interest is $t_{244} :=$ 2016-01-01. The set of index scores belonging to day t_n is thus: $GT(t_n) = \{G_{s,i}, s \in \{1:974\}, i \in \{1:244\} | (s+i = n+1)\}$. For each day of interest ($n \in \{244:974\}$) the set size $m(t_n)$ of relevant GT index scores for that day equals 244. An averaged GT index score for a day t_n, $\overline{P}(t_n)$ is then calculated by averaging the calibrated values in $GT(t_n)$, where the calibration factor w_s for an index score from sample s still has to be determined. This leads to the following formula:

$$\begin{aligned} \overline{P}(t_n) &= \frac{1}{244} \sum_{(s,i)|(s+i=n+1)} G_{s,i} \cdot w_s \\ &= \frac{1}{244} \sum_{s=1}^{974} \sum_{i=1}^{244} G_{s,i} \cdot w_s \cdot \delta_{(s+i),(n+1)} \end{aligned} \tag{1}$$

The calibration factor is determined by the overlap between two consecutive GT samples. There is an overlap of 243 days and the sum of all index scores of those overlapping days is compared. The first GT query GT_1 is taken as a reference. The second series is then calibrated to the first series by a factor C_2 for which:

$$\sum_{i=2}^{244} G_{1,i} = C_2 \sum_{j=1}^{243} G_{2,j} \Rightarrow C_2 := \frac{\sum_{i=2}^{244} G_{1,i}}{\sum_{j=1}^{243} G_{2,j}} \tag{2}$$

In general, C_k is determined by the following equation for $k \geq 2$ (for $k = 1$, $C_1 := 1$):

$$C_k := \frac{\sum_{i=2}^{244} G_{k-1,i}}{\sum_{j=1}^{243} G_{k,j}} \tag{3}$$

By induction, each query GT_s is calibrated to the first query GT_1 by the calibration factor:

$$w_s = \prod_{k=1}^{s} C_k. \tag{4}$$

In total, this gives:

$$\overline{P}(t_n) = \frac{1}{244} \sum_{s=1}^{974} \sum_{i=1}^{244} G_{s,i} \cdot \prod_{k=1}^{s} C_k \cdot \delta_{(s+i),(n+1)} \tag{5}$$