One Shot, One Hit?

Investigating, Predicting and Simulating First Contact in the European Social Survey

by

Nicholas Heck-Groβek

Department of Sociology

City, University of London

Submitted for examination in April 2021 for the degree of

Doctor of Philosophy

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AUC	Area Under (ROC) Curve
CAPI	Computer Assisted Personal Interview
CATI	Computer Assisted Telephone Interview
CRONOS	ESS Cross National Online Survey
ERIC	European Research Infrastructure Consortium
ESS	European Social Survey
(ESS-)FOPSim	European Social Survey Fieldwork Optimisation Simulation
FMS	Fieldwork Monitoring System
FN	False-Positive
FP	False-Negative
GLM or GLMNET	Lasso and Elastic-Net Regularized Generalised Linear Models
INSEE	Institut national de la statistique et des études économiques
KFCV	k-Fold Cross-Validation
KNN	k-Nearest Neighbours
KPI	Key Performance Indicator
LOGIT	Binary Logistic Regression (not log-odds)
LOOCV	Leave-One-Out Cross-Validation
MICE	Multiple Imputations by Chained Equations
ML	y
MSA	Metropolitan Statistical Areas
MSE	Mean-Squared-Error
NIR	No Information Rate
NSD	Norwegian Centre for Research Data
NZV	Near-Zero Variance
OLS	Ordinary Least Squares
PCA	Principal Components Analysis
PoC	Proof-of-Concept
RF	Random Forest
ROC	Receiver-Operator-Curve
SVM	Support Vector Machine
TN	True-Negative
TP	True-Positive
TSE	Total Survey Error
XGB	Extreme Gradient Boosting

Acknowledgements

‘Focus on the project, do not get distracted, avoid larger changes in your life’, was the most common advice I got when I started this adventure. Following this advice was not as easy as I thought. Since then, I moved places four times, became a father for the first time, got married, found a job, left a job, found a new job in my desired career, successfully handled multiple side projects, became a father for the second time, started reconstructing a house and now juggling everyday challenges during a once-in-a-century pandemic. There is a good chance that I will remember these years as the most exciting ones in my life. Without the support of many of my friends and family I would not have been able to finish this thesis.

I thank Dr ‘meta-cop’ Eric Harrison and Dr Sally Stares for being my supervisors. I did not forget how the three of us ended up being a team and I am still grateful for your offer in supervising my project. Thank you for your support over the years and keeping me on track.

Without the help of you, Nabil Manzoor, I would not have made it very far. You were my mentor and friend. I will never forget our coffees on Chatsworth Road market in the mornings and painting the town red in the evenings. Thank you for always being a generous host when I need a place to stay and giving me the comfort of having a second home in the UK.

The saying ‘it’s the friends we make along the way’ has never been more applicable than for you Sonila Dardha. Even if everything went wrong, it would have still been worth all efforts just for making friends with you. We were team SoNic from day one, inseparable until today even when we are hundreds of kilometres apart. I miss working with you for hours and heading off to the pub after a long day of work. It still fascinates me how well we work as a team. Thank you for your support over the years! I am most thankful for having you as a friend.

I know that my in-laws, Martina and Holger and all their children, will be as relieved after submission as me. You were always there for me, and especially the kids, when I struggled handling everything at once. Thank you for treating me like one of your own children and a brother.

My mum uses the phrase, ‘if it scares you, it’s probably worth a try’ and were I not raised with this mindset, I would probably not be where I am now. To my beloved parents and my admired brother: no words can express how thankful I am for your support, giving me courage, making me believe in myself, and bringing me back on my knees when I am down.

Two little children and one unbelievably strong woman are probably the three people who supported me the most. Your laughter and joy are what brightens every single day for me. You, Anja, were the one who supported me in leaving, never doubted this journey, were my shield against the outside world and put aside your own desires and aspirations to be my crutch whenever I needed you. I can only imagine the extra burden you carried. Thank you for everything you do and making me a better person.

During my last years in school, I worked late-shifts as a cashier in a local supermarket. One evening I spotted the ‘PhD’ on a young woman’s credit card. That night I decided that I too wanted to have these three letters on my credit card one day. Ten years later, I am almost there. Wishful thinking has brought me here.

‘Et elle montrait naïvement ses quatre épines. Puis elle ajouta:

›› Ne traîne pas comme ça, c’est agaçant. Tu as décidé de partir. Va-t’en.‹‹

Car elle ne voulait pas qu’il la vît pleurer. C’était un fleur tellement orgueilleuse…’

(Antoine de Saint-Exupéry)

- to the power of wishful thinking -

Abstract

Face-to-face surveys are still considered the gold standard in academic survey research despite being challenging in practical implementation and therefore resource terms. This thesis aims to contribute to a better understanding of fieldwork processes and increase their efficiency. In early phases of the survey fieldwork, researchers and practitioners face a difficult challenge: from a methodological perspective, it is necessary to establish contact with each sampled unit to avoid nonresponse or sample selection biases. At the same time, reducing the number of contact attempts as much as possible is desirable to decrease fieldwork costs. A successful contact at the first attempt would be the solution to this challenge. Consequently, knowing more about how to predict first contact success would be valuable both theoretically and practically. Previous research has already identified various 'correlates of contact', primarily using standard statistical techniques like logistic regressions. This thesis starts by providing a review of the literature to synthesise and clarify the state of research. Then both standard statistical as well as machine learning techniques are applied to investigate how these correlates play out in three large European countries (United Kingdom, Germany, France) engaged in the European Social Survey. This provides not only a set of empirical results for the three countries, but also a methodological comparison of the different statistical techniques, which can inform future survey fieldwork in practice and show that using machine learning approaches to answer survey methods research questions is feasible. The last empirical chapter presents a prototype for a simulation approach, which successfully tailors the attributes of a contact attempt to the

characteristics of a potential respondent to maximise the probability for a successful first contact.

PART I

'If we knew when each household was at home, then the interviewers could visit the housing unit at those times and contact would be established on the first visit in each case.'

Groves & Couper, 1998, p. 80.

1. Introduction

The mode of face-to-face interviewing is considered the gold standard in survey research. Despite its high costs and the potential interviewer effects which are associated with face-to-face surveys, the advantages of a standardised interviewer-administered mode outweigh the disadvantages (Groves 2009). Data quality from face-to-face surveys is deemed to be higher in comparison to other modes, mostly due to the role of an interviewer in establishing cooperation and increasing response rates, administering the questionnaire and helping respondents deal with difficult survey items (Groves 2009). Yet, in recent decades, face-to-face surveys have suffered from continuously declining response rates (Groves and Couper 1998, p. 159; de Leeuw and de Heer, 2002; Luiten et al. 2020; Schnell 1997, p.11; Stedman et al. 2019), which have further contributed to increases in fieldwork costs. Thus, finding ways to improve the efficiency and effectiveness of face-to-face survey operations is crucial, especially in the light of competing cheaper modes such as web surveys and process or user-generated data, which are increasing in popularity.

The aims of this thesis in its broadest terms are to contribute to a better understanding of fieldwork procedures, to explore approaches that could improve fieldwork efficiency and effectiveness and, as a result, develop a method that could reduce fieldwork costs to make face-to-face surveys more competitive. Thanks to the complexity of the face-to-face survey life cycle (Groves 2009), there is a multitude of possible starting points and the need for improvement is large in many of the processes associated with conducting surveys. For example, amendments to sampling techniques or to the use of computer-assisted interviewing could contribute to better survey efficiency. This thesis focusses on the fieldwork operation stage and specifically on one particular phase: the first contact attempts an interviewer makes to try to recruit a potential respondent.

Regardless of the exact characteristics of fieldwork operations, to conduct the face-to-face survey an interviewer must be assigned to reach out to a sampled unit. In the context of this thesis a ‘unit’ can be any target of the sampling process, whether it is a household or an individual. The interviewer will visit the sampled address or the address which corresponds to the sampled unit, depending on the sampling procedure. Establishing contact is a prerequisite for establishing cooperation and thus crucial for the survey procedure. The outcomes of a contact can be the immediate commitment to the interview, a soft or hard refusal, the scheduling of the interview at a later point in time or a non-contact.

When a unit refuses to participate in the survey, some information regarding the possible respondent and the contact status of the target unit is obtained as long as contact is established. This is also true even if only a unit other than the target was contactable, who was able to give at least some information. When a target refused participation or took part in the survey, the case can be closed, and the interviewer can move on to the next target unit. If the interviewer schedules an appointment with the target unit at a later point in time, the interviewer – and the fieldwork agency – can plan their next working days accordingly. In all these cases the visits yielded a satisfying fieldwork disposition code rather an unsatisfactory non-contact which is similar to a missing value.

The worst-case scenario, on the other hand, occurs when no contact with anyone can be established at the target address. In such a case, an interviewer visits the address in vain, and does not know whether the target unit would participate in the survey or not. Consequently, the unsuccessful contact attempt not only results in an unsuccessful interview, but it also fails to provide any meaningful information about the unit. The only feature associated with this contact attempt is cost. The interviewer needed time to visit the unit's address, which they could not use to reach out to another unit, who might have been easier to contact.

For data quality reasons, as explained later, the interviewer must revisit the address and try to make contact on at least a second attempt. The first and all potential follow-up contact attempts are time-intensive and are thus associated with direct and opportunity costs for fieldwork agencies. Reducing costs, while maintaining data quality, is a central goal for any fieldwork agency to maximise efficiency. If fieldwork agencies had more information on a unit's contactability, they could tailor their interviewers' visits and plan fieldwork more cost-efficiently, with the aim to successfully contact as many units on the first attempt as possible to avoid unnecessary travel and hourly payments. Even if agencies knew that the probability of contacting some units was rather low, they could at least reduce their uncertainty and estimate future costs more precisely. Of course, it is also possible to optimise contact attempts other than the first, but the benefit from optimising fieldwork is higher the earlier a contact can be obtained, which is why this thesis focuses on the first contact attempt.

Reducing cost is particularly important because conducting large scale surveys is extremely expensive (Stoop et al. 2010, p. 81-82). Although the exact costs vary from study to study, depending on the design and requirements of the study, Kreuter and Müller (2015) give an example of estimated savings of 1,800 US dollars for a 6,000 cases telephone survey by implementing a better contact strategy based on paradata. They point out that – despite other complications – even higher reductions can be expected if strategies like this are successfully implemented in face-to-face surveys which are more expensive than the telephone mode. The benefits from this optimisation are obvious and savings could either be used to reduce the overall survey costs or to fund other important steps in the survey process. Reducing any survey costs related to unnecessary travel and follow-up attempts is in the interest of all stakeholders involved in a survey research project.

One research project that tries to address and reduce fieldwork costs through monitoring is the European Social Survey (ESS). In the ESS Round 9 a 'Fieldwork Management System' (FMS) was introduced to monitor the effectiveness of all fieldwork related processes and outcomes alongside providing recommendations if fieldwork outcomes did not meet the expected quality requirements. To achieve this, fieldwork related Key Performance Indicators (KPIs), including non-contact rates, were made available to relevant stakeholders almost in real-time during the survey fieldwork process to provide timely and consistent information about the fieldwork progress (Douhou et al. 2018, p. 1-3). Additionally, the weekly reported KPIs, for example the observed non-contact rates, were compared against weekly projections of estimated fieldwork outcomes, which were generated by every national coordinator prior to the fieldwork phase for planning purposes. The rationale for this procedure is getting a better understanding of the ongoing fieldwork process, evaluating the progress and counteracting any fieldwork obstacles as early as possible. For the comparisons between projections and observed fieldwork outcomes, the authors of the 'ESS Round 9 Guidelines on Fieldwork Monitoring' ask the national coordinators of the ESS in the respective participating countries to make 'realistic' (Douhou et al. 2018, p. 1) projections and base them on previous fieldwork experiences of the ESS if possible, or on factors like the size of the interviewer workforce, interviewer workload and known issues in contacting units. However, no information was given on *how* to exactly project the fieldwork outcomes like the non-contact rate. Hence, methods for a precise projection – or prediction – might be of value for national survey coordinators.

In addition to the FMS, more extensive recommendations are made available to participating countries of the ESS to reduce nonresponse. The 'Guidelines for Enhancing Response Rates and Minimising Nonresponse Bias' cover the training of interviewers, the need for discussing the ESS specific contact procedures with the executing fieldwork agencies, interviewer workforce, interviewer workload (limited to 48 interviews per interviewer), interviewer payment (hourly rate versus per-completed-interview, while only paying for completed interviews is not recommended), minimising refusals and other forms of nonresponse through appropriate doorstep behaviour and refusal avoidance techniques, as well as minimising non-contact (Stoop et al. 2018, p. 7-17).

The recommendations address the general requirement to meet an overall non-contact rate of at most three percent, but they are also relevant for the specific aim of increasing first contact successes, which is the focus of this thesis. On the organisational level, the guidelines state that the fieldwork period can last for up to four months and with a minimum of at least one month, as longer fieldwork periods increase the chances of reducing overall non-contactability. Additionally, national coordinators are expected to plan the fieldwork with regards to their national context, for example avoiding the holiday season in countries where holiday seasons have been shown to be problematic for establishing contact. On an interviewer level, interviewers are expected to make at least four in-person attempts to reach an assigned unit before that case is abandoned over a timespan of at least 14 days, at different days of the week and times of day, one of which must be at the weekend and one in the evening hours (Stoop et al. 2018, p. 10). These complex requirements show how complicated a contact process can become and the large potential for cost reductions if a method was found that maximises the probability of contacting a unit at first attempt and consequentially avoids the need for at least up to three more revisits. On the positive end, fieldwork agencies would have an estimate of whether a unit will be contacted or not before the interviewer even arrives at the address. However, agencies must not exclude from the sample those units with a small probability of contact, for methodological reasons outlined later. If fieldwork agencies understood fieldwork processes even better or had predictions about contact success at hand, they would be able to calculate fieldwork costs accordingly.

The thesis is structured into two parts. The first part reviews the relevant literature, presents the state of research, and introduces the reader to the data and methods used in the later chapters. The second part of the thesis contains the analytical chapters, which deal with investigating the first contact attempt by using standard statistical techniques but also machine learning methods and ultimately by developing a simulation framework – named 'FOPSim' – to find the optimal set of contact attempt features for each individual based on their attributes. Reminiscent of the 'Tailored Design Method' (Dillman et al. 2014) and 'Tailored Fieldwork Design' (Luiten and Schouten 2013) this set of optimised contact attempt features will be called 'Tailored First Contact'. To begin the investigation of whether such a simulation approach is feasible and useful or not, the relevance of the topic is first established in Chapter 2. The substantial focus of the topic will be framed in its survey methodological context in Section 2.1. This section will sensitise for the challenge of balancing the financial burden, which ultimately translates to as few contact attempts as possible, and the need for as many contact attempts as possible to guarantee high data quality.

Section 2.2 reviews the literature on contact procedures and shows how difficult it is to relate contact success back to a single correlate or predictor or even to a distinct set of measures, given that various possible factors are entangled with each other and vary across countries. While it might be helpful for survey researchers to sub-divide these possible correlates into area, household, and unit characteristics, which influence first contact attempt success, it is ultimately most important to understand how to optimise the contact efficiency given these characteristics. For instance, fieldwork practitioners cannot change the composition of a household, or whether a unit has children or not, but they can assign the best interviewer at the right time on the right day to maximise contact probability. There has already been a lot of research focusing on trying to find the best possible contact strategy (e.g. Durrant et al. 2011). Most investigations on call-scheduling use some form of logistic regression models to identify at which time units are most likely to be contacted and what influences this outcome. This choice of modelling is useful since the outcome of interest is typically binary coded as 'contact' versus 'no contact'. Additionally, most of the more recent studies have applied multilevel modelling techniques to deal with the complex hierarchical structure of contact data where household units may be nested within interviewers or areas. To illustrate, multilevel discrete hazard rates were used in Stoop (2005) and Durrant et al. (2011), multilevel logistic regression was used by Blohm et al. (2006) and Wagner (2013), multilevel cross-classified regression was performed in Lipps (2016), and Wang et al. (2005) as well as Durrant and Steele (2009) applied multilevel multinomial logistic regression techniques. The literature also shows other types of analysis such as coefficient decomposition using logit models applied in Blom (2012) and Markov Chains used by Greenberg and Stokes (1990). While it seems that logistic regression models have gained some prominence in this area of research, this thesis makes use of machine learning methods that are less commonly applied in the social sciences and compares the results to a logistic regression model in the later chapters. The fact that the first contact attempt has rarely been the principal focus of any study merely serves to underline the relevance of this thesis.

The European Social Survey provides the data for the analyses of this thesis and will be fully introduced in Chapter 3 alongside explanations of the data pre-processing procedures, inclusion criteria and statistical methods used in the analyses of the later chapters. The insights from the literature review will be used in Section 3.3 to operationalise the derived concepts for the analysis. A link to a GitLab repository, which contains all annotated code to reproduce the analyses, is also provided in this chapter as well as technical details for contextualisation.

In addition to making a substantial contribution to the field of survey methods research by analysing first contact attempts, a second objective of this thesis is to trial methods from the field of data science and machine learning to investigate whether they are useful approaches to answer questions of survey practice and figure out their feasibility for a simulation approach to predict future fieldwork outcomes. Concepts, techniques, terminology, and methods from the field of data science often sound unfamiliar to scholars of the social sciences. Sections 3.8 and 3.9 will show that the main differences between social and data scientists really lie in terminology and the scope of their analyses. However, data science tools and methods do not diverge a lot from traditional social statistics. As a matter of fact, it will be shown that machine learning can be seen as a useful addition to traditional statistics, and primarily exploits advances in computational processing power and storage capabilities. On closer inspection both quantitative empirical social science and data science are well-founded on similar computational concepts. The aim of Section 3.9 is to familiarise researchers from a social science background with the data science concepts as well as to introduce the methods for the subsequent chapters.

The second part of this thesis is dedicated to the empirical analyses of first contact success. Chapter 4 focuses on the univariate and bivariate analysis of the first contact attempt in the ESS Round 9 in three countries – United Kingdom, Germany, and France – which were selected as example countries for all analyses. The analysis will investigate four high-level research questions to evaluate whether the findings from the literature review find support. It will be shown that most of the findings from previous research are supported. Most importantly, large cross-country differences will be highlighted which emphasise the importance of contextualisation for contact success.

Based on machine learning methods, Chapter 5 predicts the first contact success of units in the ESS Round 9. It will be shown that while a prediction of contact success is feasible, the predictive performance varies a lot between the three countries and depends heavily on the deployed input dataset. The best performing models show an Area Under the Curve (AUC) value ranging from 57.6 to 61.0.

To cope with the perceived downsides of limited sample sizes observed in Chapter 5, Chapter 6 makes use of pooled ESS data from Rounds 1 to 9. In a first step the associations between the independent variables and a first contact attempt success and their potential changes over the nine survey rounds are investigated. It turns out that many of the associations are not clearly stable over time. The predictive performance of the algorithms indicates that the expected benefits of the increases in sample sizes do not unambiguously translate to improvements in the predictive performance of the algorithms, although they are known to work better the with larger sample sizes. When leveraging the sample sizes from the pooled data the best performing algorithms achieve an AUC value ranging from 58.1 to 61.1. These minor improvements lead to the question whether the additional effort is outweighed by the limited performance gains. On the other hand, when comparing all models, it seems that the larger sample sizes stabilise the predictions and lead to increases especially in models which performed particularly poor with limited sample sizes.

These analyses culminate in an attempt to develop a ‘Fieldwork Optimisation Simulation’ in Chapter 7. By utilising the best performing algorithms from Chapter 6 for each country, the most optimal contact strategy for a given unit is determined by simulating all possible combinations of a set of contact related attributes like the time of the day or interviewer gender. While the approach is successful overall, it is shown that there are country differences, and that the operational capability needs to be discussed in the context of the predictive performance of the underlying algorithms.

Lastly, the thesis concludes with an overarching summary of all chapters, consolidating findings from the analyses alongside a discussion of their implications for survey practice and research practitioners.

2. Contactability in Survey Research

The previous chapter introduced the importance of reducing survey costs in general and the costs associated with the first contact attempt in particular by increasing the success of first contact attempts as much as possible. This chapter will serve two purposes: first, the trade-off challenge between economic efficiency and ensuring survey quality will be framed in its survey methodological context of the Total Survey Error (TSE) framework in Section 2.1. Secondly, the literature on the correlates of contact will be summarised in Section 2.2 to lay the foundation for the analyses in the later chapters of this thesis.

To start putting the first contact attempt into its survey methodological context, it is important to point out that the field of survey methodology has already dedicated ample attention to unit nonresponse and its affiliated problems for data quality over the last decades, (see, for example, Felderer, Kirchner, and Kreuter 2019; Peress 2010; Schnell 1997; Stanley et al. 2020; Stoop 2005). A large and ever-growing body of literature deals with nonresponse, especially examining the declining response rates (see, for example, Krejčí 2007; de Leeuw and de Heer, 2002; Luiten, Hox, and de Leeuw 2020; Schnell 1997; Stoop et al. 2010) despite all efforts made by survey methodologists to manage fieldwork better, train interviewers or incentivise respondents. As a result, with declining response rates, the costs of conducting surveys has increased significantly (Stoop 2005, p. 3). The first section of this chapter will now familiarise the reader with the fundamental concepts of survey methodology and the importance of the TSE framework for survey research as well as how nonresponse and contactability fit in the survey life cycle.

2.1. The Total Survey Error Framework

Dealing with various potential errors in a survey environment and preventing or reducing them is central to the role of the survey methodologist to establish and maintain high data quality.

Poor data quality, on the other hand, arises when there is a deviation between the respondent's answer and their actual, unobserved 'true' value (Esser 1997) – referred to as 'survey error' – or if an estimated survey value ($\hat{\theta}$) differs from the true population value (θ), which is referred to as 'survey (in)accuracy' or 'measurement error' (Noack 2015, p. 36). Identifying, tackling, and preventing threats to good data quality are paramount to obtaining accurate and reliable survey results. To conceptualise all potential risk factors for survey quality, TSE emerges as the main framework of the survey life cycle (Brown 1967; Weisberg 2005). The TSE framework is a multifaceted structure that aims to formalise the understanding of all possible error sources that might influence the survey quality at different stages of its development. The main idea of TSE is to have one single frame that summarises the two main elements of error sources: measurement, largely from the questionnaire, respondent or the interviewers, and representation, largely from the sampling frame, sampling technique or other processing errors. Although the TSE can be presented as a formula, it is not meant to be calculable, but to be seen more as a theoretical framework depicting potential influencing factors on survey quality (Biemer 2010).

The 'Mean Squared Error' (MSE) is a concept which underlies the TSE framework. The idea of the MSE is to split all potential errors into two groups: biases and variances. Bias (notated as $B(\hat{\theta})$) is described as the systematic deviation of a survey estimate from its true population value (Groves 2009, p. 52). The reasons for a systematic deviation could be many, but as an example, consider a female interviewer, who is unintentionally yet systematically, influencing the respondent's answer to a question on violence against women simply because of her sex. Answers, which are altered to acquiesce to the interviewer or to avoid their judgement are prompted from a so-called 'social desirability bias'. In this example, some participants might not give their true answer should they tolerate violence against women, but instead they edit their answer to report the socially desirable statement of condemning it, just because of the interviewer's sex. Systematic response deviations due to interviewer characteristics are referred to as 'interviewer biases'.

Variances (notated as $V(\hat{\theta})$), on the other hand, can also be introduced in numerous ways and describe the influence of random errors on survey estimates (Groves 2009, p. 53). Consider a health survey in which interviewers need to take blood pressure tests with the help of a sphygmomanometer but were not told where exactly to place the meter for measurement. Some interviewers might always choose the correct position for the sphygmomanometer; some might always, and thus systematically, use the wrong arm, which would introduce measurement bias. However, yet another group of interviewers might use the left arm on one day, the right arm the other day, take measures from the correct spot every now and then, but keep positioning the blood pressure monitor too high or too low on other days. If there is no pattern or relationship between this volatile behaviour and other (observed) characteristics, this can be assumed to be an unsystematic or randomly introduced error, and thus, be classified as so-called 'measurement variance'.

The MSE conceptualises the different biases and variances in the following formula, which shows the MSE as the aggregate of the squared bias and variance estimators:

$$MSE = \hat{\theta} = \left[E\left(\hat{\theta} - \theta\right)\right]^2 = B\left(\hat{\theta}\right)^2 + V\left(\hat{\theta}\right) \tag{1}$$

Although it is not feasible, the MSE can theoretically be calculated for every estimator in a dataset. However, to do this, error-free data for validation purposes, for example, from registers is needed. Unfortunately, registers or other validation options are seldom available. Hence, the MSE is almost never calculated in practice, which means that it rather remains in a theoretical form to simply visualise and conceptualise potential error sources.

Equation 1 is of crucial importance in the field of survey methodology as it is a holistic representation of survey quality. It can be said with no doubt that most survey methodological efforts aim to minimise this aggregate. A reduction can be realised if both $B\left(\hat{\theta}\right)^2$ (bias) and $V\left(\hat{\theta}\right)$ (variance) approach zero. If the expectancy value of an estimator does not differ from the true population parameter – i.e., if the squared bias $B(\hat{\theta})^2$of this very parameter is zero – this parameter is referred to as being *unbiased*. Additionally, an estimator can be defined as *precise*, if the variance of this estimator $V(\hat{\theta})$ is small. If both conditions are true, i.e., bias and variance are simultaneously small, the estimator is called *accurate* (Biemer and Lyberg 2003, p. 818; Noack 2015, p. 36). Building on this overarching framework, bias and variance can be split into subcategories. According to Schnell (2012, p. 387) there are five different sources of bias: specification bias, coverage bias, nonresponse bias, measurement bias and data processing bias, as well as three subcategories of variance: sampling variance, measurement variance and data processing variance. With these more detailed subcategories in mind, Equation 1 can be rewritten in the following notation (Noack 2015, p. 36):

$$MSE = \left(B_{Spec} + B_{Cov} + B_{NR} + B_{Meas} + B_{DP}\right)^2 + \left(Var_{Samp} + Var_{Meas} + Var_{DP}\right) \qquad (2)$$

As mentioned earlier, each component of the MSE can have severe impacts on survey data quality. Preventing any negative influence from these sources on any survey estimate given a restricted financial budget and limited time is paramount for establishing best practices in survey methods (Biemer and Lyberg 2003, p. 821; Noack 2015). Survey researchers are particularly interested in achieving accurate estimates that have a small bias and variance; hence it becomes their task to deal with and minimise factors that can potentially threaten the accuracy and precision of their estimates. The reader can find more information on the TSE framework in Weisberg (2005).

2.1.1. Nonresponse and Nonresponse Bias

Throughout this thesis, the analyses investigate concepts that refer to both nonresponse and the related nonresponse bias. Therefore, this section focuses on these two concepts in more detail. Although other components of the TSE framework are equally important for estimator accuracy, they are not in the scope of this thesis and will not be discussed further.

Nonresponse can generally refer to two different types: on the one hand it describes the phenomenon that a unit refuses to answer to a specific survey question (for example, to the income question) or to certain parts of the questionnaire (for example, to a set of sensitive questions about sexual activity). This is typically referred to as *item nonresponse*. On the other hand, a unit cannot take part or refuses participation in a survey, which is referred to as *unit nonresponse*. Item nonresponse can also occur because units forget to answer questions, due to a lack of motivation or poorly designed filters. Item nonresponse can introduce severe problems for data quality as it can significantly lower the number of observations available for the analysis, which will get obvious in the analyses of Chapter 5. Consequently, it has been in the focus of a lot of research and is one of the key fields for both survey researchers and survey practitioners. Many strategies have been developed to prevent item nonresponse with the help of better questionnaire design (Dillman et al. 2014) or the development of complex statistical imputation methods, for example using the ‘Multiple Imputation by Chained Equations’ (MICE) approaches (Azur et al. 2011; Groves 2009, p. 45-46).

For the purposes of this thesis, the so-called *unit nonresponse* is of more interest. Unit nonresponse means that a unit did not participate in a survey at all. Although it is easy to imagine various reasons why a unit does not want or is unable to participate, unit nonresponse can generally be attributed to one of three reasons (Schnell 2012, p.156). First, some units might be *unable to respond*. They might for example be temporarily or chronically ill, deaf, dyslexic or even illiterate or might not speak the language the survey is conducted in. Second, despite all efforts, some units *cannot be contacted*. This can be due to a wrong address, vacant property or simply the fact that the respondent is not at home for the duration of the fieldwork period or at least at the times when the interviewer tries to contact the unit. Third, as most scientific surveys conducted by universities or research institutes are voluntary – in contrast to, for example, the German or UK Census, which are legally compulsory – units make use of their right to *refuse participation*.

Looking at differences between nonrespondents and respondents is the centrepiece of nonresponse research, dating back to Rose (1959) who discussed the differences between respondents and nonrespondents in their social participation. Since the early days of survey research, scholars have been worried that units who participate in a survey differ systematically and significantly from those units who do not participate, and that this deviation impacts on, or more precisely *biases*, the survey estimates.

A common belief, that needs to be ruled out, is that a low response rate – and thus a high nonresponse rate – inevitably leads to poor data quality. In fact, nonresponse alone does not necessarily constitute a problem for a survey. Surveys suffering from high nonresponse ‘only’ need to cope with small(er) sample sizes and hence increased variances of the estimates, which result in less precise findings if the nonresponse is *random* and does not occur systematically. However, nonresponse becomes a greater threat to data quality if units that did not respond differ systematically in their (unobserved) answers from units that did respond. In other words: if people, who did not answer the survey – or parts of it in the case of item nonresponse – are *systematically* different from units who did, nonresponse bias might affect the accuracy of survey estimates (Bethlehem et al. 2011, p. 3). This is the case of nonresponse that cannot be ignored. Formally, nonresponse bias can be expressed as:

$$Bias(y_R) = y_n + \frac{NR}{n} * (y_R - y_{NR}) \quad (3)$$

With $\frac{NR}{n}$ as the proportion of nonrespondents out of all sampled units n (the so-called 'nonresponse rate'), y_R as a statistical value (for example, a mean) for a specific variable from the respondents R, y_n as a theoretical value for the same variable of all sampled cases *n*, and y_{NR} as an unobserved statistical value for the same variable of all nonrespondents *NR*. Nonresponse bias can then be defined as the product of the nonresponse rate and the difference between the respondents' mean and the nonrespondents' mean of the same variable. The common misconception that high nonresponse (or low response rates) alone inevitably leads to poor survey data quality can be dismantled by looking at this equation. The nonresponse rate $\frac{NR}{n}$ is not problematic in itself as long as the deviation in respondents' and nonrespondents' answers $(y_R - y_{NR})$ approaches or is equal to zero, i.e., there is no difference between them (Groves 2006, p. 648, 2009, p. 59, 189; Groves and Couper 1998). The nonresponse rate is one component of the error but on its own introduces neither nonresponse error nor bias. This does not mean that nonresponse can be overlooked, nor that actions which are taken to reduce the likelihood of nonresponse should be decreased. Of course, preferably researchers want to have information on all eligible units, and thus, it is in the interest of data accuracy to include every eligible unit. Following the argumentation so far and considering the different reasons nonresponse occurs as discussed earlier, Equation 3 can be respecified as follows:

$$Bias(y_R) = y_n + \left(\frac{UN}{n} * (y_R - y_{UN})\right) + \left(\frac{NC}{n} * (y_R - y_{NC})\right) + \left(\frac{RF}{n} * (y_R - y_{RF})\right) \quad (4)$$

Equation 4 breaks down Equation 3 into the different reasons unit nonresponse manifests, and accounts for the situation that units who were unable to participate (*UN)* are different from units who were not contactable (*NC)*, and from units who refused participation (*RF)* (Groves 2004, p. 134). Equation 4 contains a rate for each specific dropout group: $\frac{UN}{n}$ is the share of nonrespondents who were unable to participate, $\frac{NC}{n}$ is the non-contact rate and $\frac{RF}{n}$ is the refusal rate, relative to all units in the sample n. Besides the raw rate, Equation 4 also features each specific group's deviation from the respondents' values with $y_R - y_{UN}$ being the deviation of unable units from respondents, $y_R - y_{NC}$ being the deviation of non-contactable units from respondents and $y_R - y_{RF}$ being the deviation of refusing units from respondents (Groves 2004, p. 134).

This distinction is important when working with nonresponse problems (Groves and Couper 1998, p. 80; Lynn and Clarke 2002) as they need to be approached differently. Different actions can be taken to tackle the specific causes of nonresponse and different sub-fields of survey research dedicate efforts in examining the specific reasons for nonresponse (Stoop 2005, p. 50). Survey instruments and procedures are being designed, tested, revised, adapted and improved to include even those units who would typically not participate in a survey for various reasons. For example, interviewers are trained in so-called ‘refusal conversion’ methods to convince people who initially refuse participation to answer the survey. In the context of this thesis unit nonresponse attributed to non-contactability is of central importance. Thus, the next section will go into more detail why non-contactability can pose a threat for survey data quality.

2.1.2. (Non-)Contactability and First Contact Success

The first contact with a unit is of primary interest in this thesis. This section, thus, elaborates on non-contactability in more detail. Because of a potential nonresponse bias, the assumption that there is no difference between contacted and non-contacted units is risky to say the least. However, the assumption is also usually not testable since a lot of information on non-contacted units is naturally missing. As it cannot be assumed that there is no difference between contacted and non-contacted units, scientific surveys make immense efforts and dedicate large financial resources to ensure that everyone who should be contacted and should be part of the survey will be contacted.

Making sure that no group of units is systematically left out of the contact process is crucial for data quality. This is part of what is commonly referred to as ‘representation’ in the TSE framework or ‘representativity’ or ‘representativeness’ in everyday language – although including all necessary groups does not cover the whole list of requirements to make a sample 'representative'[1]. If a survey fails to include all groups which exist in the general population, the sample can suffer from the so-called 'sample selection' or 'sample composition' bias. This could be, for example, because no contact could be established for a specific group of units who share certain characteristics. This means that the final survey population excludes some units systematically and that it is not equal to the frame or general population (Kalton 1983, p. 5-7). The consequences of sample selection biases jeopardise the quality of the survey as a whole and may render the survey results useless.

For a successful face-to-face contact at first attempt, it is logically necessary that both the interviewer and the unit are present at the same time and at the same location – in this context usually the respondent's home. Different at-home times of units and different working schedules (or contact times) of interviewers create certain fruitful time windows for each unit-interviewer pair, which make contacts possible. Due a multitude of factors affecting at-home patterns as well as interviewers' working times, meeting these time windows on the first visit can be exceedingly difficult and is subject to chance. Because not everyone can be contacted at the first attempt, most fieldwork agencies follow-up uncontacted units a specific number of times before they assign non-contact as the unit's final disposition code.

From a methodological perspective, it is highly desirable to establish contact, regardless of the resources required to do so because of the risk of a potential nonresponse bias.

[1] Furthermore, it is important to be specific about the characteristics a sample is representative for. If a happiness survey gets a representative balance of socio-demographic characteristics, it might for example still systematically undersample people suffering from depression. Consequently, *the* representative survey does not exist and the word 'representativeness' makes no sense without specifying for which characteristics it is representative.

From an economic point of view, however, this is typically not feasible as these follow-ups can come with large expenditures of time and/or financial resources for the interviewers or the fieldwork agency. The interviewers need to travel to all previously non-contacted units, who may be geographically distant, several times and might expect payment from the survey agency for all these contact attempts. This leads to a challenge between increasing contact attempts to avoid disadvantageous sample composition and to maintain data quality on the one hand, and financial constraints to stay within a limited budget, on the other hand. Unfortunately, budget is typically a scarce resource in most research projects and social science projects are no exception. Consequently, this challenge can be framed as an optimisation problem to ensure that the largest possible amount of contact attempts is made for a fixed budget.

In the best-case scenario, contact would be established successfully right at the first attempt. Unfortunately, this is not necessarily the case. For the National Survey of Health and Stress, which was fielded in 48 United States between September 1990 and February 1991 (Groves and Couper 1998, p. 69) the authors report that about 50 percent of the respondents have been contacted in the first contact attempt (Groves and Couper 1998, p. 102). Another 20% were contacted at the second attempt, while further 20% were still not contacted after the third attempt. Thus, in roughly half of the cases interviewers needed to visit the target address at least two or more times to avoid the risk of sample selection biases. In a comparison of 67 surveys fielded in Germany between 1984 and 1993, Schnell (1997, p. 217) found that roughly 10 percent of the reasons for missing data in these surveys can be attributed to non-contactability. As discussed, non-contactability alone does not necessarily pose a threat to data quality. However, if units who are contacted more easily differ systematically from those who are not and insufficient efforts are made to include hard to contact units, sample selection biases might occur. Research shows that a sampled population gets more 'urban' with increasing numbers of follow-up calls (Groves and Couper 1998, p. 109). In other words, after only one or two follow-ups, a sample is lacking a significant number of units who live in urban areas. If one stopped contacting after only a few follow-up calls, the survey sample would not necessarily resemble the population it aims to refer to in terms of urban/rural distribution. Similar results were found for the age of a respondent, their household composition, education level and employment status, indicating that more contact attempts lead to a better fit between sample and general population compositions (Stoop 2005, p. 168; Vicente 2017). However, the exact number of follow-up visits that are needed to obtain a reasonable sample composition remains ambiguous in the literature. Stoop (2005, p. 184) finds efficiency gains in survey quality when up to six contact attempts are made spread across times of the day and days of the week. Sturgis et al. (2017) found only small differences after three contact attempts and negligible differences in sample composition after five contact attempts. In the most extreme case, Wang (2005) found no meaningful differences after only one follow-up attempt. Despite their differences, these studies share the implicit underlying finding that more resources which needed to be spent on follow-up visits were saved, if more units were contactable at the first attempt.

The remarks and findings from this section suggest that exploring the relationship between contactability and other (survey) characteristics are interesting for survey methodologists for various reasons. Therefore, the literature review in the following Section 2.2 covers the research on reasons for (non-)contactability and correlates of contact.

2.2. Research on Correlates of Contact

The field of survey methodology has been trying to understand the reasons for establishing contact for at least the last thirty years. A large body of research focuses on identifying correlates of the contact process and finding ways to improve survey fieldwork with so-called ‘adaptive’, ‘tailored’ or ‘responsive’ designs (Groves and Heeringa 2006; Luiten and Schouten 2013). Two quotes from the literature describe the purpose of these designs and emphasise the importance of the inevitable trade-off between quality and costs:

> *‘Whether designs are altered during fieldwork, or whether they are tailored to specific subgroups before fieldwork begins, what these approaches have in common is a differential fieldwork strategy, aimed at minimizing nonresponse bias and survey costs, while trying to maintain survey response at a level that is necessary for precise survey estimates’* (Luiten and Schouten 2013, p. 170)

and:

> *‘The main aims of such designs are to increase response rates (for example by prioritizing high response propensity groups), to decrease nonresponse bias (for example by prioritizing low response propensity groups), to reduce survey costs (for example by limiting the number of calls), or any combination of the previous, sometimes competing goals’* (Vandenplas et al. 2017, p. 660)

This section summarises important findings, which relate to the contactability of a unit in general and when possible, also to the first contact attempt in particular. Despite the large amount of research conducted in this space, the literature shows that many of the findings are ambiguous since the probability of a successful contact is connected with a variety of interactions between individual factors.

The following sub-sections summarise the research on correlates of contact. The term ‘correlates’ is used here to describe all the factors which are associated with contactability and the contact process. Groves and Couper (1998, p. 264) state that acknowledging them as correlates and not as causes is an important distinction to make, even though a causal connection appears to be very likely in some cases. Nevertheless, testing for causal inference is impossible in observational studies and would require controlled randomised trials (Hox et al. 2006). The empirical chapters in this dissertation assume some underlying causal connection between the correlates and contactability and try to predict the binary outcome (‘contact’ or ‘non-contact’) of the first contact attempt. This means causality is assumed as part of model fitting. In other words, the ‘correlates’ of contact are used as independent variables, factors or predictors that influence whether an interviewer manages to successfully contact a respondent in the first attempt.

Although the focus of the literature review lies on establishing contact, it is important to also introduce a brief view of the research on survey participation. Contact processes and survey participation are closely linked to each other, but participation is evidently the ultimate goal of conducting a survey, which is why research on participation enjoys a lot of methodological attention. To begin with, some of the most fundamental research on establishing contact or participation has been summarised both theoretically and empirically by Groves and Couper (1998), who introduced and discussed extensively the concept of survey participation. In addition, a sophisticated multilevel analysis of fieldwork process, interviewer and respondents' effects on participation can be found in Lipps (2009), while Lipps (2008), Durrant et al. (2010) and Jäckle et al. (2013) focus on specific and very detailed interviewer influences on participation in a telephone and (several) face-to-face surveys, respectively. A comparison of six UK surveys in Durrant and Steele (2009) shows that some predictors of contact and participation have opposite effects from one another. Additionally, some predictors are unstable and change their influence depending on which survey they look at. They find out further that linking administrative data adds less precision to estimates than anticipated. Overall, the literature on survey participation is just as ample as the one on establishing contact. However, participation is beyond the scope of this thesis to cover in detail, but it is worth mentioning that research on contactability and research on participation share a lot of similar predictors, which attests that the two fields are closely linked to another.

Besides their work on participation, Groves and Couper (1998) also dedicate their research to contactability. Their book *Nonresponse in Household Surveys* builds the foundation for a large body of literature conducted in the subsequent years following its publication. Perhaps one of the most important claims in this book is that the importance of fundamentally distinguishing between non-contacts and refusals cannot be underestimated, as these groups of units can be differently related to survey variables and thus can affect the survey estimates differently (Groves and Couper 1998, p. 80). Consequently, the authors argue that they must be considered as two different kinds of problems: one is establishing contact, the subsequent is establishing cooperation and participation. The authors further divide the sphere of potential correlates of contactability into four groups: social environmental area indicators of at-home patterns[2], household-level correlates, interviewer-level correlates and call-level influences (Groves and Couper 1998, p. 84). Most of the research in the subsequent years has followed this categorisation and the next sections summarise the research utilising the same categorisation. It is worth mentioning, however, that a distinct allocation to only one specific group proves to be difficult for some factors/variables, due to interactions between them. Table 1 on page 48 gives an overview of the important studies which constitute the literature review.

2.2.1. Social Environmental Area Indicators of At-Home Patterns

Potential units can only be contacted at their address if they are at home at the time the interviewer seeks to visit them. Overall, the literature presents a main finding that reasons for being at home can cluster both spatially and temporally. Regarding spatial clustering, inhabitants of large towns seem to spend more time travelling to and from their workplace than inhabitants of urban areas, which makes them less available at their dwelling, all else equal (Robinson and Godbey 1997). Stoop (2005, p. 66) reports evidence from the Netherlands where especially city-dwellers, single people and units with higher education participate in cultural events outside of their homes, which makes them harder to reach.

Groves and Couper (1998) expect inhabitants of large cities to spend more time on grocery shopping or entertainment options, like cultural events, outside of their homes. They suggest that these differences in behaviour can be linked to the population size of a sampling

[2] A more modern term might be ’socio-geographical data’.

point or the urban size of an area. Overall, they find statistically significant evidence that large metropolitan statistical areas (MSAs) have lower first contact success rates than small MSAs (Groves and Couper 1998, p. 86). Additionally, comparisons between low-density areas and high-density areas show that units are reached more easily the less dense an area is (Campanelli et al. 1997) and thus, the more urban an area gets, the more contacts are needed to successfully contact a household (Durrant et al. 2011; Stoop 2005, p. 168). Furthermore, areas with the highest density also showed a higher number of units that were ultimately non-contactable compared to the low-density areas (Groves and Couper 1998, p. 87). Bivariate findings investigating population density and urbanicity hold in several multivariate analyses as they appear to significantly impact contact success even after controlling for other potential confounding factors (Groves and Couper 1998, p. 107; Hox et al. 2006; Luiten and Schouten 2013; Stoop 2005, p. 169). Findings from comparative research, however, show that these relationships differ by country. For example, negative effects of urbanicity on contactability (i.e., positive effects for rural areas on early contact success) were found in Finland and Ireland, while in the United Kingdom, Belgium, Greece and Spain no differences were found and in Portugal, results even showed a positive influence of urbanicity on contactability (Blom 2012).

Despite the undeniable importance of urbanicity-measurements, researchers suggest that they are likely confounded with other indicators proven to be influential for contact success. Research shows that areas with a high amount of multi-unit houses (i.e., towns or cities) show lower contact success compared to areas with a lower amount of multi-unit houses (i.e., rural areas) (Campanelli et al. 1997; Stoop 2005, p. 55). Areas with a higher amount of owner-occupied properties (i.e. rural areas) mark higher first contact success rates compared to areas with a lower amount of owner-occupied properties (i.e., towns or cities) (Groves and Couper 1998, p. 87). Other potential confounding factors are the marital status of units (fewer married units in highly populated areas) as well as the amount of crime in an area (more crime in high-density areas) (Groves and Couper 1998).

Next, aggregates of individual or household characteristics can also be observed at area levels and can influence at-home patterns. As mentioned, Groves and Couper (1998) present evidence that the higher the share of owner-occupied homes in an area, the higher the first contact success rate is. They argue that tenants are on average less wealthy and younger than house owners and their lifestyle might involve more out-of-home activities, which make them harder to contact (Groves and Couper 1998, p. 85f). Furthermore, Dennis et al. (1999) found that the median income of a geographical area is a successful predictor for contact rates. Early work of Juster and Stafford (1985) finds more out-of-home time for ethnic minority groups. This is supported by other research, which finds a negative relationship between the proportion of ethnic minority groups in an area and the predicted contactability of units in the area and for households with members of a minority group in particular (Brick et al. 1996; Groves and Couper 1998, p. 86; Luiten and Schouten 2013). Interestingly, Wang et al. (2005) show contradictory results with higher first contact probability for areas with higher proportions of minority groups. Contrasting both these findings, Blohm et al. (2006) find no significant influence of belonging to a minority group on contactability whatsoever. Overall, the literature of the influence of the proportion of units who belong to a minority group in an area on contactability is as mixed as the research on social environmental area indicators in general.

2.2.2. Household-Level Correlates

The household level incorporates all the characteristics that are mutually shared by the units within a household. It also refers to the physical characteristics of a building, which are also shared by all the members who live in this building, for example, the state of repair of a multi-dwelling house, which is occupied by several independent households.

Among the most investigated factors associated with contactability at the household level are physical access impediments like locked gates, doormen or intercom systems, which can prevent an interviewer from entering a building easily or at all. Research finds that the presence of such access impediments reduces the likelihood of the first contact success (Blakely and Snyder 1997; Durrant et al. 2011; Groves and Couper 1998, p. 88; Hox et al. 2006; Lipps and Benson 2005; Wang et al. 2005). Groves and Couper (1998, p. 89) show that contact success rates in the United States were much higher for units with no physical access impediments such as bars on windows, security doors, crime watch or security system signs, locked entrances etc., compared to those buildings that had these security features. In addition, they found that at the end of the fieldwork period, seven percent of households with access impediments remained uncontactable, while only two percent of those without access impediments remained ultimately uncontactable. In another study covering the United States, Cunningham et al. (2005) found that 17 percent of all dwelling units had some form of controlled access feature. In seven percent it was an intercom/buzzer, six percent had physical barriers like gates, doors or locks and four percent were guarded in person by security or doorpersons. While these barriers are an impediment in the context of face-to-face surveys, telephone surveys come with their very own set of impediments that can reduce contact probability. Among those are (automated) call rejection, voicemail systems or simply a busy line during the telephone call attempt (Vicente 2017).

Although findings of several national studies support the negative effect of access impediments on interviewers being able to establish contact and the increase in nonresponse, the effects again vary between countries. Blom (2012) shows that intercoms reduce contact success in Portugal but increase contact rates in Spain. The author argues that interviewers in different countries might have a different perception of what an intercom represents: for some, an intercom may be perceived as an access impediment that is typically connected to households of low socio-economic status. For others, an intercom may be perceived as a direct access to a unit's home and considered to relate to wealthier, newer and maintained households or buildings. Durrant and Steele (2009) do not find significant influences from any of these interviewer access impediments on contact probability. Overall, the findings on the presence and significance of the influence of access barriers on contact success remains context-sensitive in the literature and needs further investigation despite the already high number of publications on this topic.

Relationships have also been found between the quality of a housing unit and contact success, suggesting lower contact success for low housing quality (Durrant et al. 2011; Durrant and Steele 2009; Hox et al. 2006; Lipps and Benson 2005). In the UK and Finland bad housing quality was found to be statistically significant and negatively related to contact propensity while no statistically significant association was found for Belgium and Portugal (Blom 2012). However, while the condition of construction appeared to be negatively associated with contact success in some studies (Stoop 2005, p. 175), it did not show an effect in other studies (Groves and Couper 1998, p. 88).

Literature further suggests that household composition also has an association with contact success. A single unit household can only be contacted if this individual is at home, while a multi-unit household, on the other hand, can be reached even if only one of the household members is present during the contact attempt(s). As a result, the single unit households have lower contact propensities. Findings from previous research show that there are large differences in contact success for single unit households compared to households with more adults. Generally, households with a larger number of adults were more likely to be contacted (Durrant et al. 2011; de Leeuw and de Heer, 2002; Lipps and Benson 2005). Other research shows that large households are contacted in fewer calls compared to single unit households, which need significantly more contact attempts to reach the same contact success rate (Groves and Couper 1998, p. 91; Stoop 2005, p. 174). Durrant and Steele (2009) find that there is an additional difference between multi-unit households and couple-households with the latter having an even higher contact rate than the former. They theorise that multi-unit households often consist of students or young professionals who might share their dwelling, but, apart from that, have independent lifestyles from one another, which resemble the lifestyle of singles more than the one of a family. The multi-unit households, thus, have higher contact rates than single unit households due to the larger number of adults living there, but do not reach the high contact probability of couple-households.

Furthermore, literature shows that households with children or elderly people also have a higher chance of being contacted. Households with children are often found to be more easily contacted compared to single unit households or those without children (Durrant et al. 2011; Durrant and Steele 2009; Groves and Couper 1998, p. 91; Luiten and Schouten 2013; Lynn and Clarke 2002; Stoop 2005, p. 174). Similarly, households with elderly people have a higher contact rate, especially emphasising here the higher first contact rate, than households without elderly people (Durrant and Steele 2009; Groves and Couper 1998, p. 91; Hox et al. 2006; Luiten and Schouten 2013). In the same vein, studies show that ‘young' units aged between 18 and 44 are harder to contact on the first attempt (Vicente 2017; Weidman 2010). Findings on the number of adults per household, the presence of the elderly as well as the presence of children were robust to inclusion of possible confounders (Groves and Couper 1998, p. 107).

Units who actively participate in the labour force and, more broadly, units with a higher socio-economic status, are harder to contact than those of a lower socio-economic status (Durrant and Steele 2009; Goyder 1987, p. 84; Smith 1983; Stoop 2005, p. 174). Consequently, households with units who are not part of the labour force, retirees or (temporarily) unemployed units, show higher contact success propensity (Stoop 2005, p. 66). Finally, some investigations also focus on sex differences in contact success. Groves and Couper (Groves and Couper 1998, p. 136) argue that women are still more frequently involved in care duties than men which is why they might be easier to contact. Stoop et al. (2010, p. 14) also relate sex effects to labour-market differences and argue that in some European countries, different female work-patterns lead to women being at home more often, and thus, contact success with them is more likely.

2.2.3. Interviewer-Level Correlates

Correlates on the interviewer level summarise all effects that relate to the interviewer, who might affect the contact success of a potential respondent. Computer-Assisted Personal Interviewing (CAPI) surveys rely heavily on the performance of interviewers and this dependence cannot be disregarded. Interviewers carry out crucial tasks, ranging from managing the gross sample of list of addresses, contacting units, establishing participation, and maintaining cooperation of units throughout the survey interview. Given this importance of interviewers, a large body of research is dedicated to interviewer effects (West and Blom 2017). Interviewer influences not only start at the stage of persuading units to participate, but even earlier when trying to establish contact. As West and Blom (2017) display in their review of interviewer effects, these influences can have large impacts on the success of a survey, including on unit nonresponse, item nonresponse and response.

Overall, findings for interviewer effects on contactability are as complex as the previous findings on social environmental area and household-level correlates. Some previous inquiries find no relationship between interviewer experience and contactability (Durrant et al. 2011; Groves and Couper 1998, p. 95), while other research finds that more experienced interviewers have a lower rate of contacts than less experienced interviewers (Wang et al. 2005). Also, other research looks into self-reported attitudes of interviewers towards their work motivation. Interviewers who describe themselves as trying to convert refusals instead of accepting refusals have a slightly – but not significantly – higher contact rate (Groves and Couper 1998, p. 95). The same applies to interviewers who have more confidence in their refusal conversion skills: the higher their confidence, the higher the contact success rate (Groves and Couper 1998, p. 95). Further, interviewers' attitudes towards work and their expectations regarding their success is also related to the contact rates they get (Singer et al. 1983). Other research found that interviewers who are good at establishing cooperation also tend to have higher contact rates (Durrant and Steele 2009; O'Muircheartaigh and Campanelli 1999). Interviewer workload on the other hand was found to be negatively correlated with contact rates (Blom 2012).

Being able to establish a phone call with a unit before visiting in person was also found to increase the in-person contact probability in some studies (Lipps and Benson 2005; Schnell 1997, p. 220). However, in her international comparison, Blom (2012) only found support for this finding in Finland, where a large proportion of units were contacted via phone beforehand. Stoop (2005, p. 169) concludes that the absence of a telephone number is a practical impediment rather than a general indicator of being hard to reach face-to-face. Similarly, mode of first contact was investigated in Blohm et al. (2006) finding no significant influence on first contact attempt probability while Durrant et al. (2011) indicate a worse in-person contact performance for interviewers using initial phone contacts.

Other internationally comparative findings (Blom 2012) indicate that there are differences in contact propensities for interviewer sex, education, work experience as well as work behaviour e.g., interviewers leaving a note at the respondent's house. However, such results vary in their direction and strength between countries. In other studies, it was also shown that older interviewers, those with higher education and those working in the afternoon hours are more successful in establishing contact compared to younger, less educated and those interviewers only working on weekends (Durrant et al. 2011; Hox et al. 2006). Durrant et al. (2011) also found evidence that differences in pay schemes lead to different contact success rates, with better paid interviewers having higher contact rates.

2.2.4. Call-Level Correlates

Call-level correlates are associated with all potential characteristics around the contact attempts themselves. For at least thirty years, the most investigated factor related to contact success has been the time of the contact attempt. As seen previously, at-home patterns generate time windows in which units are contactable and interviewers must meet these time windows to successfully establish contact. The research on call timing has brought up many studies, focusing primarily on improving fieldwork processes to find out how to meet these call windows. However, Wagner (2013) concludes that despite all efforts, research on improving call schedules have not yielded the desired practical effects and, considering the level of complexity of this topic, this is not surprising.

Interviewers working on face-to-face surveys might not necessarily be full-time workers and often have other jobs or at least other daily duties. Thus, it is not uncommon that they are self-employed or working independently as contractors. This also means that they mostly decide working hours themselves within a suggested time frame. Arranging these time frames is important to reduce respondent burden, for example seeking to recruit respondents while they may be working at home or engaged in household or caring duties, and to ensure ethical fieldwork procedures, e.g., not contacting households too early in the morning or too late in the evening. For example, it might be efficient to contact units very late at night as they are very likely to be at home during this time, but disturbing potential respondents in such late hours would also be unethical as it might for example cause distress.

Survey methodologists have investigated the optimal call scheduling for both modes of telephone and face-to-face surveys and have focused on variations of days of the week, times of the day or more general seasonal fluctuations, to find out at which times units are most likely contactable (Cunningham et al. 2003; Kulka and Weeks 1988; Massey 1996; Weeks et al. 1980, 1987). Evidence found across a multitude of studies suggests that the majority of research is very clear about the positive relationship of weekday afternoon and evening calls as well as weekend calls on contact success (Campanelli et al. 1997; Durrant et al. 2011; Durrant and Steele 2009; Lipps and Benson 2005; Purdon et al. 1999; Stoop 2005, p. 160f; Vicente 2017; Wagner 2013; Wang et al. 2005; Weeks et al. 1987). These findings are robust to the inclusion of other predictors of contact success, showing that morning and midday calls on weekdays have a negative impact on the contact probability.

However, research has also brought to light that optimal call times can vary for specific subgroups quite substantially. A detailed analysis by Durrant et al. (2011) investigated optimal call schedules for different household characteristics attributed to different at-home patterns and found variations in the optimal call timing based on these characteristics. They found that residents of houses are easier to contact in the afternoon on any day of the week, while residents of flats are more easily contacted in the evenings or weekday mornings. This again might be confounded with the presumption that units living in houses might be older on average or might have children, while units in flats might be younger units without children. Units of houses that are in fair or bad conditions of repair are more easily contacted on Thursday to Sunday mornings compared to any other time. This might also be due to an interviewer strategy of predominantly contacting units in such areas in the hours of daylight, due to fear of crime. Finally, Durrant et al. (2011) found that households with children are more easily contacted on weekday afternoons whereas households without children are most successfully contacted in the evening hours. Thus, the optimal call timing is likely to be subject to different household characteristics.

When there was a previous unsuccessful contact, weekday evenings and weekends appear to yield the highest contact probability for a follow-up contact (Durrant et al. 2011; Weeks et al. 1987). However, the literature discusses the effectiveness of altering the contact times of subsequent calls if previous ones were unsuccessful. Some research finds no benefits from altering contact times (Brick et al. 1996; Groves and Couper 1998, p. 192; Vicente 2017), while other studies suggest that calling at the same time again increases contact probability (Greenberg and Stokes 1990; Kulka and Weeks 1988) while a third branch of research finds positive effects of altering contact times to improve the likelihood of contact (Durrant et al. 2011).

Additionally, studies in a longitudinal setting often make use of data from previous waves (Lipps 2012; Trappmann et al. 2015). Findings from these panel analyses suggest contacting units of longitudinal studies at those days of the week and times of the day when they were initially interviewed in their first successful survey wave (Laurie et al. 1999; Lipps 2012). Interestingly, Kreuter and Müller (2015) bring a contradictory result to light: they tested whether call scheduling at previously successful times improved contact success, and only found support for Lipps (2012), who suggested the call scheduling method, in the observational part of their study but not in their experiments. They found an even stronger effect on cooperation when units were contacted at the same time of the last waves' successful interview window compared to the last waves' successful first contact time in the observational study. Overall, however, they conclude that the findings from the observational study may be due to selection biases in the sample composition instead of being clear effects of the call timing. In other analysis of panel data Trappmann et al. (2015) showed that important changes in a unit's life, like becoming full-time employed, losing a job or moving the household, negatively relates to the contact probability for a subsequent wave. Conversely, Blom (2012) found a stable effect for all analysed seven countries (UK, Belgium, Finland, Greece, Ireland, Portugal, Spain) with the amount of calls being negatively related to the final contact success: the more calls were made, the less likely the unit was finally contacted. A similar result was also found by Durrant et al. (2011). These results underline the importance of the first contact attempt. Overall, such evidence not only shows that the exact effect or effect direction remains context specific.

While a large body of literature deems evening and weekend calls to produce better contactability success, Blom (2012) presents evidence that makes the discussion even more complex. After decomposing coefficients, Blom (2012) finds results contrary to the majority of the literature, indicating no clear positive influence of evening and weekend calls on contact success between countries. In line with Stoop (2005, p. 54) the author explains that the positive effect of evening and weekend calls might be confounded with unobserved interviewer contact strategies, who favour specific time windows, only because they know from their personal experience in the job or even have local knowledge that reaching out to units is more likely at these specific times. Thus, they artificially increase the success rate at specific times. In such a scenario, success would not be directly related to the fact that the unit was called on a weekday morning, but rather because the interviewer had heuristic knowledge about the best contact time. How exactly the interviewer gains this knowledge remains unclear, but it is nonetheless important to mention that interviewer strategies might confound correlations on contact times. Scenarios like this might also explain unexpected findings like in Campanelli et al. (1997), who found that experienced interviewers tend to call during the daytime preferably before switching to evening calls although the latter are considered to be more effective. These experienced interviewers might have developed their own interviewer strategy which remains unobservable to the research team. Also, such a finding underlines again the expectation that correlates of contact may work together rather than independently.

These studies (Blom 2012; Campanelli, Sturgis, and Purdon 1997; Stoop 2005) make it apparent how interweaved the interviewer and call levels are. This comes as no surprise since the interviewer chooses the time of contact, and thus, is the ‘administrator’ of the call level. Also, this means that the interplay between interviewers and call correlates must be investigated in more detail. A major challenge is that, in contrast to CATI studies, interviewers in face-to-face studies are often not randomly assigned to their contact units (West and Blom 2017). This means that *any* found effect can possibly be an artefact or selection effect due to an unobserved interviewer strategy (Hox et al. 2006). Interviewers are predominantly working independently and choose their working times mostly themselves and fieldwork guidelines recommend this approach to maintain interviewer motivation and fieldwork feasibility. Furthermore, research highlights that interviewers, who plan their schedule of work themselves, have higher contact rates, as they contact respondents more likely at the weekends and evenings (Lievesley 1983). All these contact strategies are beneficial when they increase fieldwork outcomes. However, they might introduce biases, which in most cases cannot be accounted for as they may be unknown to the researchers.

For example, one of the interviewer strategies may be daytime visits, which are more likely a practice in those areas that have a higher perceived crime rate. Interviewers may try to avoid visits in the hours of darkness in such sampling points because they fear victimisation. Unfortunately, if daytime contacts are less successful, there is a higher chance that units from these areas remain ultimately uncontacted, and thus underrepresented in the survey, if they get contact attempts at these unpreferable times of the day (Durrant and Steele 2009; Lepkowski et al. 2010). Stoop (2005) describes that these interviewer strategies introduce a severe problem for 'hard-to-reach' units, because if they do not contact the units at optimal times, these units become artificially hard-to-reach only as a result of an interviewer's strategy and not necessarily because they are actually so. She notes that this also works in the opposite direction: if interviewers know that specific areas have a high proportion of working units, they might start enforcing more evening calls there and thus artificially improving (first) contact rates (Stoop 2005, p. 53-55).

Another example of how interviewers can positively influence contact rates comes from Durrant et al. (2011), who looked at interviewers' behaviour and whether they leave a note behind after an unsuccessful contact attempt. Unsurprisingly, they found that those interviewers who left a notification, showed higher contact rates at the next call. Further, Wagner (2013) conducted sophisticated experiments to find out more about interviewer behaviour and found the dissatisfying result that interviewers did not follow the fieldwork protocol in a face-to-face survey because it proved to be highly impractical. He notes that due to feasibility, interviewers work by area and thus try to contact all units in an area once they are there, rather than driving from that area to another one and then back and forth only to reach out to units at their optimal times. He concludes that, even if research finds optimal call times for households, the next step is to introduce a trip-planning optimisation for interviewers, which again, can result in a trade-off between interviewer assignment feasibility and optimal call-scheduling. Overall, interviewer strategies influence the first contact attempt success quite heavily and generally make it a challenging correlate for contactability, as researchers cannot account for any interviewer strategies if no data on these are available – which is mostly the case.

The before-mentioned analyses on contact timing rely mostly on interviewer contact-sheet data. In most surveys, interviewers fill in a form about their contact process themselves; this contact data can usually only be generated automatically in specific modes like in a CATI survey wherein process-generated data is much easier to record. Biemer et al. (2010) analyse the effect of falsely reported fieldwork outcomes in contact-sheet data and conclude that they bias nonresponse-correction-models notably. This finding raises awareness of the possibility of fake contact protocols which researchers might be dealing with, which in turn limit the credibility of the data the researcher can use. Unfortunately, researchers most commonly have no choice but to assume that their data is legitimate if there is no clear evidence of fraud. An exhaustive literature review on the quality of paradata can be found in West and Sinibaldi (2013).

Although it is challenging to summarise the findings on the call-level correlates, most studies found that overall, weekday afternoons and evenings as well as weekends were the most effective contact times. Nevertheless, optimal contact times are different for different household (and by extension, also individual) characteristics and thus, these can be complex relationships. Furthermore, when looking at contact success in general rather than the first contact specifically, there is no clear evidence about how to alter contact time attempts or how to deal with contactability over the span of a fieldwork period or even multiple survey waves. Finally, the role of the interviewer and their employed strategies during fieldwork is central to understanding contactability. Further research is needed to examine how these strategies could be incorporated into contact designs.

In this chapter, the challenge to find the right balance between the number of necessary contact attempts to contribute to the overall data quality and economic efficiency has been framed in the survey methodological context. It was emphasised that nonresponse bias as well as sample selection biases might threaten the data quality when units, who should be included into the survey, are not included. It was further shown that one group of unit nonresponse can be attributed to units who are non-contactable. In the literature review of this chapter, it has been examined that the contactability of a unit depends on various factors that can be related to the social environmental area of the unit, the household, the interviewer or the contact attempt itself. Moreover, it was shown that associations are not only interweaved with one another but are also highly context specific and vary between countries. Following Blom (2012), differences in fieldwork outcomes between countries may occur due to divergences in survey implementation, population characteristics and different effects of population characteristics on contact probabilities in a specific country due to cultural variations. Despite, or even because, their high contextuality, the findings from the literature review are of major importance and interest for the remainder of this thesis as they lay the foundation for the analyses in the later chapters.

Before these analyses however, the next chapter first presents the data and methodology for these analyses. One key component of Chapter 3 is the operationalisation of those variables that were derived from the literature review.

Author	Year	Title	Country	Mode
Biemer	2010	*Total Survey Error: Design, Implementation, and Evaluation*	Multiple	Multiple
Blakely, Snyder	1997	*Fortress America: gated communities in the United States*	USA	-
Blom	2012	*Explaining cross-country differences in survey contact rates: application of decomposition methods: Cross-country Differences in Survey Contact Rates*	UK, Belgium, Finland, Greece, Ireland, Portugal, Spain	f2f
Brick	1996	*Outcomes of a Calling Protocol in a Telephone Survey*	USA	Telephone
Campanelli, Sturgis, Purdon	1997	*Can you hear me knocking? and investigation into the impact of interviewers on survey response rates*	UK	f2f
Cunningham, Flicker, Murphy, Aldworth, Myers, Kennet	2005	*Incidence and Impact of Controlled Access Situations on Nonresponse*	USA	f2f
Cunningham, Martin,Brick	2003	*An Experiment in Call Scheduling*	USA	Telephone
de Leeuw, de Heer	2002	*Trends in household survey non-response. A longitudinal and international perspective*	Australia, Belgium, Canada, Germany, Denmark, Finland, France, Hungary, Italy, The Netherlands, Poland, Sweden, Slovenia, Spain, UK, USA	Multiple
Dennis, Saulsberry, Battaglia, Rodén	1999	*Analysis of Call Patterns in a Large Random-Digit-Dialing Survey: The National Immunization Survey*	USA	Telephone
Durrant, D'Arrigo, Steele	2011	*Using paradata to predict best times of contact, conditioning on household and interviewer influences*	UK	f2f
Durrant, Steele	2009	*Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys*	UK	f2f
Goyder	1987	*The silent minority: nonrespondents on sample surveys*	Multiple	Multiple
Groves	2009	*Survey Methodology*	Multiple	Multiple

Groves, Couper	1998	*Nonresponse in household interview surveys*	USA	f2f
Hox, Blohm, Koch	2006	*The Influence of Interviewers' Contact Behavior on the Contact and Cooperation Rate in Face-to-Face Household Surveys*	Germany	f2f
Juster, Stafford	1985	*Time, goods, and well-being*	USA	Multiple
Kreuter, Müller	2012	*A Note on Improving Process Efficiency in Panel Surveys with Paradata*	Germany	CATI/CAPI
Kulka, Weeks	1988	*Toward the development of optimal calling protocols for telephone surveys: a conditional probabilities approach*	USA	Telephone
Laurie, Smith, Scott	1999	*Strategies for reducing nonresponse in a longitudinal panel survey*	UK	f2f
Lepkowski, Mosher, Davis, Groves Van Hoewyk	2010	*The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey*	USA	f2f
Lievesley	1983	*Reducing Unit Non-response in Interview Surveys*	UK	f2f
Lipps	2012	*A Note on Improving Contact Times in Panel Surveys*	Switzerland	f2f
Lipps, Benson	2005	*Cross-national contact strategies.*	Austria, Belgium, Denmark, France, Germany, Greece, Italy, Netherlands, Spain, Sweden, Switzerland,	f2f
Luiten, Schouten	2013	*Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction*	The Netherlands	Multiple
Lynn, Clarke	2002	*Separating refusal bias and non-contact bias: evidence from UK national surveys*	UK	f2f
Massey	1996	*Optimum calling patterns for random digit dialed telephone surveys*	USA	Telephone
O'Muircheartaigh, Campanelli	1999	*A multilevel exploration of the role of interviewers in survey non-response*	USA	f2f
Robinson, Godbey	1997	*Time for Life: The Surprising Ways Americans Use Their Time*	USA	Paper

Schnell	1997	*Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*	Multiple	Multiple
Singer, Frankel, Glassman	1983	*The Effect of Interviewer Characteristics and Expectations on Response*	USA	Telephone
Smith	1983	*The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey*	USA	f2f
Stoop	2005	*The Hunt for the Last Respondent*	The Netherlands	f2f
Stoop, Billiet, Koch, Fitzgerald	2010	*Improving Survey Response: Lessons learned from the European Social Survey*	30 countries from ESS4	f2f
Trappmann, Gramlich, Mosthaf	2015	*The effect of events between waves on panel attrition*	Germany	f2f
Vicente	2017	*Exploring fieldwork effects in a mobile CATI survey.*	Portugal	Telephone
Wagner	2013	*Adaptive Contact Strategies in Telephone and Face-to-Face Surveys.*	USA	Telephone/f2f
Wang, Murphy, Baxter, Aldworth	2005	*Are Two Feet in the Door Better than One? Using Process Data to Examine Interviewer Effort and Nonresponse Bias*	USA	f2f
Weeks, Jones, Folsom, Benrud	1980	*Optimal times to contact sample households*	USA	f2f
Weeks, Kulka, Pierson	1987	*Optimal Call Scheduling for a Telephone Survey*	USA	Telephone
Weidman	2010	*Do characteristics of RDD survey respondents differ according to difficulty of obtaining response?*	USA	Telephone
West, Blom	2017	*Explaining Interviewer Effects: A Research Synthesis*	Multiple	Multiple
West, Sinibaldi	2013	*The Quality of Paradata: A Literature Review*	Multiple	Multiple

Table 1: Overview of Important Studies from Literature Review

3. Data & Methodology

The previous chapters highlighted the importance of the first contact attempt in face-to-face surveys and framed the challenge between multiple contact attempts and fieldwork costs as a survey methodological problem before introducing the state of research. Chapter 3 aims to bridge the gap between the two parts of this thesis. To do so, the data and methods used in the analyses of Chapters 4 to 7 are presented in the next chapter.

Data for all analyses in this thesis comes from the European Social Survey (ESS), which was chosen as the primary dataset for various reasons covered in the next paragraphs. The European Social Survey is an academic, cross-national, repeated cross-sectional survey which collects data on individuals in the participating countries every two years. Beginning with 22 participating countries in 2002, data from the year 2018 is now available for 29 participating countries and the data release for one additional country is currently still pending in April 2021. Funding for the survey is predominantly covered by the participating countries but also supported by EU research funds. In 2013 the ESS became a European Research Infrastructure Consortium (ERIC). The ESS headquarter at City, University of London coordinates the developments made by the Core Scientific Team and the dedicated expert panels, like the Sampling and Weighting Expert Panel. Members from these panels as well as researchers from the Core Scientific Team, headquarter and the national coordinating teams make sure, that the ESS is implementing survey methodology best practices.

The ESS aims to monitor and interpret changes in attitudes and values of Europe's societies in reciprocation with Europe's' institutions, promoting advances and improvements in cross-national survey research as well as the development of European social indicators. Therefore, the ESS' objectives are twofold: On the one hand, the ESS delivers high quality substantive data about a variety of topics of relevance for the social sciences as well as the so-called 'rotating modules', which feature a specific focus questionnaire each round. On the other hand, a particularly important feature of the ESS is its focus on methodological research and innovation. The rotating modules gather information on specific domains of interest like 'Justice and Fairness in Europe' and 'Timing of Life' in ESS 9 or 'Digital Social Contacts in Work and Family Life', 'Understandings and Evaluations of Democracy' or 'COVID-19 conspiracy beliefs and government rule compliance' in the upcoming ESS 10.

The commitment to methodological excellence is the main reason why the ESS was chosen as the primary dataset for this thesis. Since its establishment almost 20 years ago, there has been a large emphasis on enabling research on interviewer behaviour and effects, nonresponse bias, data collection modes and the avoidance of measurement and translation errors. One key feature of the ESS is the availability of contact protocols and paradata for all individual sampling units regardless of whether they were contacted or not. This is also the most important reason for its selection as the primary dataset since an analysis of contact procedures is the main objective of this thesis. The comprehensive methodological standards and requirements of the ESS make it one of the social science studies with the highest scientific requirements, which aim to compare characteristics and attitudes of the majority of European populations on a variety of subjects. The rigorous methodological criteria include the commitment to strict face-to-face interviews only, random probability sampling, a target response rate of 70%, an overall non-contact rate of 3%, full transparency of methods and sophisticated documentation and as well as rigorous translation processes. It can be assumed that these rigorous standards contribute to the ESS' success and popularity which can be measured in 5,429 publications listed on Google Scholar on 11 February 2021, which averages to roughly 300 publications per year since 2003 (Malnar 2021).

One important feature is the ease of data access. Data for the European Social Survey can be downloaded for free at the website and only requires an email address for registration. This facilitates exploring and working with the data for researchers, journalists and for aspirational researchers and students in particular. Downloads are available for various datasets as well as comprehensive reports or fieldwork related documents like interviewer showcards and questionnaires. Besides the main dataset, which features the substantive interview results, data from contact protocols as well as data from interviewer questionnaires are publicly available. Main datasets can either be downloaded individually or as an integrated datafile, which combines the country datasets and already takes care of harmonisation. Most datasets are available for analyses in SAS, Stata and/or SPSS. Furthermore, the R package 'essurvey' provides R users with a powerful tool for conveniently importing ESS data into R (Cimentada 2019).

Information on the ESS as well as all related projects can be found on the comprehensive project's website (https://www.europeansocialsurvey.org) as well as in the technical reports for each round.

In this thesis, Chapter 4 and Chapter 5 focus on the analysis of contact success, making use of the latest available ESS Round 9 dataset. Round 9 was fielded in 30 countries between September 2018 and mid-2019 and features data from face-to-face interviews on a wide range of important recurring topics like media, social trust, politics, and human values as well as rotating modules on 'Justice and Fairness in Europe' and 'Timing of Life'. The sampling targeted all persons aged 15 years and older, who reside in private households, regardless of their nationality, citizenship, language[3] or legal status.

From all available 30 countries in the ESS 9, only data from the most populous European countries (United Kingdom, Germany and France) were selected as examples throughout this thesis. These countries not only represent a large proportion of the European population, but are also seen as culturally similar, which can make the results more comparable. The primary decision for the selection of these countries is that all three countries suffer from low and further declining response rates. Consequently, contributing to the understanding of this decline is paramount to maintain survey quality. An interesting detail that might have an influence on contact success is the usage of different sampling frames in the countries: while France and the United Kingdom use address-based sampling (France: Institut national de la

[3] In fact, respondents can be excluded from the survey if they do not speak the language the survey is fielded in sufficiently well. Further analysis of units with language barriers and how they differ from included units can be found in Heck-Grossek and Dardha (2020).

statistique et des études économiques (INSEE) register of dwellings, United Kingdom: Postcode Address File), register based sampling is conducted in Germany (municipality population registers).

Fieldwork in the UK was conducted by NatCen Social Research between 31 August 2018 and 22 February 2019, by the Institute for Applied Social Sciences (infas) in Germany between 29 August 2018 and 4 March 2019, and by Ipsos in France between 19 October 2018 and 1 April 2019.

The analyses of the ESS 9 data are based on the main survey dataset version 1.2 released on 31 January 2020, contact data version 1.0 released on 2 December 2019 as well as the interviewer data version 1.0 released on 31 October 2019.

While Chapter 4 and Chapter 5 focus on analysing ESS Round 9 data, Chapter 6 and 7 will make use of a pooled dataset from all published ESS Rounds 1 to 9, with two exceptions: contact data for France is not available in ESS Round 1. Since the contact data is of crucial interest in this thesis, France needed to be excluded from the analysis for this particular round. Additionally, contact data for Germany is unavailable for Round 5. According to the Norwegian Centre for Research Data, the data quality for this particular round was flawed and inconsistent and consequently excluded from the ESS 5 integrated dataset by the data holders. Even though Germany's contact data file is available as a single dataset and hence could have been linked to the remaining dataset, it has been omitted here due to data quality concerns. Since contact data is mandatory for the purpose of this analysis, all Round 5 data for Germany had to be excluded from all analyses in Chapter 6 and 7.

All necessary datasets from all rounds were downloaded on 27 March 2020 in their respective versions at that time.

3.1. Data Pre-processing

Data was either downloaded for SPSS or Stata, depending on availability, and imported into R using the 'haven' package[4]. Joins were then created manually using a combination of the unit identifier and their country label to create a unique cross-country unit identifier in each country and round. A full right join was performed between the main dataset[5] and the contact data, followed by a full left join between the previously created dataset and the interviewer data.

3.2. Inclusion and Exclusion Criteria

Observations of interest in this study are those who responded to the substantial survey as well as those who did not respond or were not even contacted. Thus, units from the datasets, regardless of their participation, are included in the analysis if they satisfied three conditions: 1. data from the contact information sheet must be available, 2. data on the outcome of the first

[4] Since, to the author's knowledge, the 'essurvey' package only simplifies the import of ESS main datasets, but not the record linkage for contact and interviewer data, this package was not used.

[5] While importing Rounds 1 to 8 was unproblematic, importing the ESS 9 version 1.2 main dataset for Stata into R for Linux with base settings provoked the error code 'Unable to convert string to the requested encoding (invalid byte sequence)'. Apparently, this error stems from an encoding difference between Windows systems with UTF-8 standard and Linux systems with Latin 1 standard. Simply extending the 'haven' command 'read_dta' by 'encoding = 'latin1'' solves this issue. This error did not occur for data from any previous ESS round.

visit attempt must be available, 3. data on the outcome of the first visit must not be ‘address invalid’ or ‘other information on unit’. Criteria 1 and 2 are mandatory when analysing which variables influence the first contact attempt success. Criterion 3 ensures that those units, who could never have been contacted because their address was wrong in the first place, are excluded from the analysis. The number of ineligible cases is an interesting finding but including these units would inflate the number of unsuccessful first contacts for non-fieldwork related reasons, which justifies their exclusion.

3.3. Operationalisation of Variables

To investigate the correlates of contact success, the variables derived from the literature in Section 2.2 were operationalised using ESS data. The operationalisation is identical throughout Chapters 4 to 7 except for the deviations for Chapter 6 explained in Section 3.4.

The outcome variable is whether a unit was contacted at the first contact attempt regardless of whether or not the contacted person was the target unit, or whether an interview took place with the respondent. The outcome variable simply distinguishes between a ‘contact’ if any contact at all was made at first contact attempt or a ‘non-contact’ if no contact to anyone at all was made at the first contact attempt. The ESS contact information data contains the variable *resulb1*, which carries information on the outcome of the first visit attempt. The outcomes of the original variable are categorised into eight levels: 1. ‘Completed interview’, 2. ‘Partial interview’, 3. ‘Contact with unidentified person’, 4. ‘Contact with respondent but no interview', 5. 'Contact with someone else than respondent', 6. 'No contact at all', 7. 'Address invalid' and 8. 'Other information on unit'. The outcomes from *resulb1* were regrouped with 1 and 2 becoming 'Completed or Partial Interview', 4 remains 'Contact with respondent but no interview', 3 and 5 become 'Contact with other than respondent', 6 remains 'No Contact at all' and 7 and 8 become missing values and are excluded from the analysis in accordance with the exclusion criteria.

In the light of this investigation, the original outcome options 1 to 5 are positive outcomes since some form of contact was established. The original outcome 6, 'No contact at all', is the negative outcome of interest in this investigation. From *resulb1* the outcome/dependent/target variable for this analysis was created to dichotomise the outcome of the first contact attempt. Outcomes 1 to 5 from the original variable were summarised as 'Successful contact' or simply 'contact' (1), while units with outcome 6 were classified as 'Unsuccessful contact' or simply 'no contact' (0).

As shown in the literature the day of contact proved to be very important for contact success. Consequently, the variable carrying information on the day of the week of the first visit (*dayv1*) was included in the analysis. This variable is available for both respondents and nonrespondents.

Besides the chosen day, the time of day also proved to be an important correlate of contact in the literature. The ESS data contains information on the interviewer reported hour as well as minute of first contact attempt (*hourv1* and *minv1*, respectively). Using these two variables a new variable was created containing the hour of the first contact attempt with minutes rounded up to the next full hour if they exceeded the half-hour mark or rounded down to the last hour if not[6]. This variable is available for both respondents and nonrespondents.

To approximate the degree of urbanicity of the unit's living area, the variable *domicil* was included from the main questionnaire in the analysis, which captures the respondent's perception of the area they live in. No changes were made to this variable. This variable is only available for respondents.

To investigate the influence of different housing types, the variable *type* was included from the contact data in the analysis. The original variable features ten levels of different housing types: 1. 'Farm', 2 'Single unit: Detached house', 3. 'Single unit: Semi-detached house', 4. 'Single unit: Terraced house', 5. 'Only housing unit in building with other purpose', 6. 'Multi-unit house: flat', 7. 'Multi-unit house: student apartments or rooms', 8. 'Multi-unit: Sheltered/retirement housing', 9. 'House-trailer or boat', 10. 'Other'. The variable was re-categorised into a more parsimonious variable: Former categories 1 to 5 become 'Single unit', categories 6 to 8 become 'Multi-unit' and categories 9 and 10 become 'Other' in the new variable. This variable is available for both respondents and nonrespondents.

The variable *physa* from the contact data was included in the analysis to investigate the interviewer's assessment of the overall physical condition of the building or house the respondent lives in. The original five-point ordinal variable reaching from 'Very bad' through

[6] In a previous analysis the exact time was used. However, the distribution of the exact time showed interviewers' heaping patterns as quarter (15 and 45 minutes), half (30 minutes) and full (60 minutes) hours were more dominant than any other minute of an hour. It was thus refrained from using the exact minute of the hour to not give a false sense of precision. Using a categorised version instead of the continuous time did not change the results significantly.

‘Satisfactory’ to ‘Very good’ was rescaled into three categories: ‘Bad or very bad’, ‘Satisfactory’ and ‘Good or very good’. This variable is available for both respondents and nonrespondents.

To assess the influence of access impediments on an interviewer’s capability to make contact, the variable *access* was included from the contact data in the analysis. The original variable captured whether there were any impediments hampering an interviewer’s access to a unit’s house or building, with the following outcomes: 1. ‘Yes, entry phone present’, 2. ‘Yes, locked gate or door present’, 3. ‘Yes, entry phone and locked gate or door present’, 4. ‘No, neither of these’. A new variable was created, and the outcomes were dichotomised into ‘No access impediments’ for outcome 4, and ‘Yes, access impediments (entry phone, locked gate/door)’ for outcomes 1 to 3. This variable is available for both respondents and nonrespondents.

To approximate whether an area can be considered impoverished and whether interviewers might perceive a higher fear of crime in these areas, the variable *littera*, capturing the interviewer’s judgement of the amount of litter and rubbish in the immediate vicinity, as well as the variable *vandaa*, capturing the interviewer’s judgement of the amount of vandalism and/or graffiti in the immediate vicinity were included from the contact data in the analysis and serve as physical signs of decay (Medway et al. 2016). The original four-point ordinal scales reaching from ‘Very large amount’ to ‘None or almost none’ were recoded into two categories: ‘None, almost none or small amount’ and ‘Large or very large amount’. Both variables are available for both respondents and nonrespondents.

To assess whether units are more likely to be contacted if a telephone number of the household is available the variable *telnum* was included from the contact data in the analysis. The variable captures whether telephone numbers for units are available or not. No information was found in the dataset or documentation regarding where the telephone numbers were taken from if they were available. This variable is available for both respondents and nonrespondents.

An approximation of the interviewer's first contact workload was created by taking the variable *intnum1*, which holds the interviewer ID for the interviewer that performed the very first contact attempt, from the contact dataset. The frequency of identical *intnum1* IDs was summed up for all observations per interviewer, to obtain a total score, which contains the total number of first contact attempts an interviewer carried out. This variable is available for both respondents and nonrespondents.

To investigate the relationship between interviewers who were more successful in establishing first contacts and those who were less successful, and the outcome variable, the total amount of successful first contact attempts was divided by the total amount of all the first contact attempts the interviewer had to process (the workload) and multiplied by 100. This variable is available for both respondents and nonrespondents. Since this variable is a function of the dependent variable of the later analyses, high multicollinearity is expected, and it will only be used for descriptive purposes.

It is crucial to note that in the ESS there are multiple variables that hold interviewer IDs to identify different interviewers. Variables *intnum1*, *intnum2* … *intnum** capture the interviewer numbers for the interviewers who carry out the first, second or subsequent contact attempts. By utilising these interviewer numbers, one can derive information like the workload as seen above. While *intnum** holds information on those interviewers who tried to establish contact, the variable *intnum* (note the absence of any wildcard/digit at the end of this variable) holds the interviewer number of the interviewer who finishes an interview. The interviewer questionnaire only captures information for and from this final interviewer (*intnum*) such as this interviewer's age and sex. Consequently, *intnum* can be used to link information on the final interviewer's age and sex from the interviewer questionnaire to the main dataset, to find out for example whether more male or female interviewers carried out complete interviews. However, sex and age are not reported for the contact interviewers, which are of primary interest in this investigation. Thus, a workaround was created to infer to the first interviewer's age and sex (identified by *intnum1*) from the final interviewer's identification number stored in *intnum*: if, within an observational unit, the *intnum* (final interviewer) is the same as *intnum1* (the first contact interviewer) then the age and sex of the final interviewer is taken from the interviewer questionnaire and linked to the interviewer of the first contact. The same interviewer who started the visits also finished the interview (*intnum1* is equal to *intnum*) in 86.5% of the participating units in the UK, in 94.4% in Germany and in 96.8% in France. This means that making use of variables that are generated through this link leads to a reduction in sample sizes of 13.5% in the UK, 5.6% in Germany and 3.2% in France when this variable is used. If the final interviewer is a different person than the first interviewer (*intnum* not equal to *intnum1*) then information related to the final interviewer (*intnum*) is *not* used for the first contact attempt interviewer (*intnum1*) and the variable value is missing. To summarise, first contact attempt interviewer's age and sex are only available for respondents and only if the interviewer ID of the final interviewer is identical to the interviewer ID of the first contact interviewer and if neither information on *intnum* nor *intnum1* is missing.

Following the same approach of combining information related to the final interviewer (link through *intnum*) and information relating to the first contact interviewer (link through *intnum1*) another variable was created to approximate how many interviews of an interviewer's first contact workload actually resulted in completed interviews. When *intnum* is added up over all observations for each interviewer, it captures how many interviews the interviewer successfully completed. Thus, by dividing the total amount of completed interviews by the total amount of first contact attempts (an interviewer's first contact workload), an approximation of the completion rate can be obtained. This completion rate serves as a proxy for interviewer success: An interviewer who has a higher share of completed interviews based on their first contact workload is labelled in this sense a more 'successful' interviewer, while an interviewer with a lower share of completed interviews relative to their first contact workload is classified as a less successful interviewer. Again, this can only be calculated when the final interviewer within an observation is the same as the initial interviewer (*intnum* == *intnum1*). Similarly as with an interviewer's age or sex, this variable is only available for respondents of the survey and only if the interviewer ID of the final interviewer is identical to the interviewer ID of the first contact interviewer and if neither information on *intnum* nor *intnum1* is missing.

Respondent's age (*agea*) and sex (*gndr*) were included from the main dataset in the analysis. These variables are only available for respondents.

ESS Rounds 1 to 8 featured a variable named *chldhm*, which carried information on whether children are living at the respondent's home or not. Unfortunately, this variable is not available in the ESS 9 dataset. Since the literature shows that this information can be a useful correlate, a workaround was conceptualised. Data from Eurostat shows that young people in Europe leave their parental household at different ages. In 2019 adolescents in the UK were 24.6 years old on average when they left their parent's home. Young adults were 23.7 years old on average in Germany and 23.6 years old on average in France when they moved out (Eurostat 2020). The variable *fcldbrn* in the ESS 9 main dataset carries information on the year when a respondent's first child was born. The variable *ycldbrn* carries information on the year when a respondent's youngest child was born. To approximate whether a respondent is living with their children, a new variable was created that makes use of a combination of this information. If a respondent's first child or youngest child is younger than or exactly as old as the rounded average age when adolescents typically leave their parental household as defined by Eurostat, this new variable indicates that a respondent lives with a child. More specifically, if a respondent's youngest child is older than 25 years in the UK or 24 years in Germany and France in the year 2019 then it is assumed that the respondent does not have a child at home. This variable is only available for respondents and only if information on the year of birth of the first and (if more than one child) youngest child are not missing.

To investigate whether the marital status of respondents has any influence on contact success or not the variable *marsts* was taken from the main dataset in the analysis. The original variable levels: 1. 'Legally married', 2. 'In a legally registered civil union', 3. 'Legally separated', 4. 'Legally divorced/civil union dissolved', 5. 'Widowed/civil partner died' and 6. 'None of these (NEVER married or in a legally registered civil union)' were recoded. Levels 1 and 2 were assigned to the new level 'Married', levels 3 and 4 became 'Separated/divorced', level 5 was reassigned to 'Widowed', and 6 became 'None of these'. This variable is only available for respondents.

The variable *mnactic* features information on the respondent's main activity in the last seven days and was included in the analysis. The original levels: 1. 'Paid work', 2. 'Education', 3. 'Unemployed, looking for job', 4. 'Unemployed, not looking for job', 5. 'Permanently sick or disabled', 6. 'Retired', 7. 'Community or military service', 8. 'Housework, looking after children, others', and 9. 'Other' were reassigned in order to build fewer levels that are more homogeneous within these groups with regards to the units' at-home patterns. Level 1 and 7 were regrouped into 'Paid work'[7], level 2 remained 'Education', levels 3 and 4 were combined into 'Unemployed', levels 5, 6, 8 and 9 remained 'Sick', 'Retired', 'Housework' and 'Other', respectively. This variable is only available for respondents.

To investigate the number of household members' possible influence on the outcome, the variable *hhmmb*, which features the total amount of people living regularly as members of the household was included from the main dataset in the analysis. This variable is only available for respondents.

[7] According to the ESS 9 core-questionnaire, community or military service 'does not apply to jobs in the military but to compulsory military and community service only' (European Social Survey 2018a, p. 52). It is assumed that at-home patterns of units in these services are like those of units in paid work regardless of whether units in military or community services get paid or not, which is why these two groups were combined.

Years spent in education (*eduyrs*) was included in the analysis from the main dataset to investigate whether they have any influence on the outcome variable. This variable is only available for respondents.

To approximate a household's wealth, the household's total net income from all sources (*hinctnta*) has been included in the analysis. The variable's original scale in ten country specific deciles was left unchanged. This variable is only available for respondents. Table 2 summarises the variable operationalisation and can serve as a lookup table for the later chapters.

Label	**New code or value range…**	**…consisting of original Code(s)**	**Original label or formula to create new variable in pseudo-SQL annotation**	**Original code, value range or explanation**	**Availability**
First contact success	0 = *No* 1 = *Yes*	6 1+2+3+4+5	resulb1	1 = Completed Interview 2 = Partial Interview 3 = Contact with unidentified person 4 = Contact with respondent but no interview 5 = Contact with someone else than respondent 6 = No contact at all 7 = Address invalid 8 = Other information on unit	Respondents and Nonrespondents
Day of the week for first visit	1 = *Monday* 2 = *Tuesday* 3 = *Wednesday* 4 = *Thursday* 5 = *Friday* 6 = *Saturday* 7 = *Sunday*	-	dayv1	1 = Monday 2 = Tuesday 3 = Wednesday 4 = Thursday 5 = Friday 6 = Saturday 7 = Sunday	Respondents and Nonrespondents
Hour of the day for first visit	0-24; minv1 rounded to next full hour at minute 31	-	hourv1 minv1	0-24 0-59	Respondents and Nonrespondents
Domicile, respondent's description	1 = *Farm or home in countryside* 2 = *Country village* 3 = *Town or small city* 4 = *Suburbs or outskirts of big city* 5 = *A big City*	-	domicil	1 = Farm or home in countryside 2 = Country village 3 = Town or small city 4 = Suburbs or outskirts of big city 5 = A big City	Respondents
Type of house respondent lives in	1 = *Single Unit* 2 = *Multi Unit* 3 = *Other*	1+2+3+4+5 6+7+8 9	type	1 = Farm 2 = Single unit: detached house	Respondents and Nonrespondents

				3 = Single unit: Semi-detached house 4 = Single unit: terraced house 5 = Only housing unit in building with other purpose 6 = Multi-unit house: flat 7 = Multi-unit house: student apartments or rooms 8 = Multi-unit: Sheltered/retirement housing 9 = House-trailer or boat 10 = Other	
Physical condition of building/house	1 = *Bad or very bad* 2 = *Satisfactory* 3 = *Good or very good*	4+5 3 1+2	physa	1 = Very good 2 = Good 3 = Satisfactory 4 = Bad 5 = Very bad	Respondents and Nonrespondents
Access impediments	0 = *No access impediments* 1 = *Yes, access impediments (entry phone, locked gate/door)*	4 1+2+3	access	1 = Yes, entry phone present 2 = Yes, locked gate or door present 3 = Yes, entry phone and locked gate or door present 4 = No, neither of these	Respondents and Nonrespondents
Litter in immediate vicinity	0 = *None, almost none or small amount* 1 = *Large or very large amount*	3+4 1+2	littera	1 = Very large amount 2 = Large amount 3 = Small amount 4 = None or almost none	Respondents and Nonrespondents
Vandalism in immediate vicinity	0 = *None, almost none or small amount* 1 = *Large or very large amount*	3+4 1+2	vandaa	1 = Very large amount 2 = Large amount 3 = Small amount	Respondents and Nonrespondents

				4 = None or almost none	
Telephone number available	0 = *No* 1 = *Yes*	-	telnum	1 = Present 2 = No Phone	Respondents and Nonrespondents
Workload	0 - ∞	-	COUNT(intnum1) GROUP BY intnum1	intnum1 holds the unique interviewer ID for the interviewer who carried out the first contact attempt	Respondents and Nonrespondents
First Contact Success Rate	0 - 1	-	((COUNT(intnum1) WHERE First Contact Success = 1)/COUNT(intnum1)) GROUP BY intnum1	intnum1 holds the unique interviewer ID for the interviewer who conducted the first visit	Respondents and Nonrespondents
Sex of first interviewer	0 = Female 1 = Male	2 1	intgndr	1 = Male 2 = Female	Available for respondents if interviewer ID of final interviewer is identical to interviewer ID of first contact attempt interviewer
Age of first interviewer	15 - ∞	-	intagea	15 - ∞	Available for respondents if interviewer ID of final interviewer is identical to interviewer ID of first contact attempt interviewer

Completion rate	0 – 1		(COUNT(intnum)/ Workload) GROUP BY intnum1	intnum holds the unique interviewer ID for the interviewer who successfully conducted the interview	Available for respondents if interviewer ID of final interviewer is identical to interviewer ID of first contact attempt interviewer
Respondent’s age	15 - ∞	-	agea	15 - ∞	Respondents
Respondent’s sex	0 = Female 1 = Male	2 1	gndr	1 = Male 2 = Female	Respondents
Children at home	0 = No 1 = Yes	-	IF country = ‘United Kingdom’ AND (2019 - fcldbrn <= 25 OR 2019 - ycldbrn <= 25) THEN 1 ELSEIF country = ‘United Kingdom’ AND 2019 – ycldbrn > 25 THEN 0 ELSEIF country = ‘Germany’ AND (2019 - fcldbrn <= 24 OR 2019 - ycldbrn <= 24) THEN 1 ELSEIF country = ‘Germany’ AND 2019 – ycldbrn > 24 THEN 0 ELSEIF country = ‘France’ AND (2019 - fcldbrn <= 24 OR 2019 - ycldbrn <= 24) THEN 1 ELSEIF country = ‘France’ AND 2019 – ycldbrn > 24 THEN 0	fcldbrn carries the information on the year when a respondent’s first child was born; ycldbrn carries the information on the year when a respondent’s youngest child was born.	Respondents

Respondent’s marital status	1 = *Married* 2 = *Separated/divorced* 3 = *Widowed* 4 = *none of these*	1+2 3+4 5 6	marsts	1 = Legally married 2 = In a legally registered civil union 3 = Legally separated 4 = Legally divorced/civil union dissolved 5 = Widowed/civil partner died 6 = None of these (never married or in a legally registered civil union)	Respondents
Main activity in the last 7 days	1 = *Paid work* 2 = *Education* 3 = *Unemployed* 4 = *Sick* 5 = *Retired* 6 = *Housework* 7 = *Other*	1+7 2 3+4 5 6 8 9	mnactic	1 = Paid work 2 = Education 3 = Unemployed, looking for job 4 = Unemployed, not looking for job 5 = Permanently sick or disabled 6 = Retired 7 = Community or military service 8 = Housework, looking after children 9 = Other	Respondents
Number of household members	0 - ∞	-	hhmmb	0 - ∞	Respondents
Years of education	0 - ∞	-	eduyrs	0 - ∞	Respondents
Household income	1 - 10	-	hinctnta	1 – 10	Respondents

Table 2: Variable Operationalisation

3.4. Deviations in Variable Operationalisation for Variables Used in Chapter 6

While the analyses in Chapters 4 and 5 utilise data from ESS Round 9, data for Chapters 6 and 7 comes from a pooled dataset of all ESS Rounds 1 to 9, which were collected over a timespan of almost 20 years. During this time fieldwork processes, operationalisation of variables, and concepts of the ESS have changed significantly and thus differences in the data and variable structure exist between ESS Rounds 1 and 9. While some ESS rounds lack some of the variables which were operationalised before entirely, other rounds might principally cover the relevant concept but use a different operationalisation or wording to survey it. Although the ESS has always committed to the highest scientific standards of survey methodology, especially the first five rounds of the ESS appear in retrospect to have been a testing field for various approaches. Between ESS 1 and 5 questioning and wording in both the main questionnaire and contact protocol changed noticeably, including answer categories and whole concepts, that were tested out in one round, dropped in the next and sometimes re-introduced in subsequent rounds. It appears that after these five rounds the ESS found the best way to ask most questions and the concepts and wording and questioning stabilised – at least for the variables under observation here. Surveying whether there were access impediments or not, was featured in the first round, but it was dropped in the second to fourth round, until it was measured again from Round 5 onwards. On the other hand, variables like sex of the interviewer were not recorded until Round 4 at all. Although the interviewers' age was included from Round 4 onwards, this was a categorical variable in Round 4, while from Round 5 onwards it was observed on a continuous scale. One striking finding is that the variable measuring whether children are living at the respondent's home or not was part of all rounds in the same format until Round 8 but was excluded from the Round 9 questionnaire.

To make sure Chapter 6 is comparable to Chapters 4 and 5 but also accounting for these variations, the operationalisation of the variables for Chapter 6 and 7 was conducted as similar to the one for Chapters 4 and 5, which is based on the ESS Round 9 data. Despite these deviations, they only concern smaller details of how subgroups were recoded. Large efforts were made to ensure that the operationalisation of all variables in all rounds resulted in comparable datasets. Outlining the full process of harmonising all variables in all nine rounds, and all countries might go into too much detail. However, the full code for the data preparation and harmonisation is available online for examination (see Section 3.7).

Table 3 gives an overview of the relevant variables and whether they were part of a specific ESS round. A dark grey coloured box indicates that the variable was covered in the respective round, while a white coloured box indicates that it was not. A box coloured in light grey indicates that the variable was covered but had a different operationalisation e.g., some other format of the question or answer categories.

Variable	ESS1	ESS2	ESS3	ESS4	ESS5	ESS6	ESS7	ESS8	ESS9
Respondent's ID	■	■	■	■	■	■	■	■	■
Result of the first visit	■	■	■	■	■	■	■	■	■
Hour of first visit	■	■	■	■	■	■	■	■	■
Minute of first visit	■	■	■	■	■	■	■	■	■
Domicile, respondent's description	■	■	■	■	■	■	■	■	■
Entry phone or locked gate/door before reaching respondent's individual door	▒	□	□	□	■	■	■	■	■
Assessment overall physical condition of building/house	□	□	□	□	■	■	■	■	■
Amount of litter and rubbish in immediate vicinity	■	■	■	■	■	■	■	■	■
Amount of vandalism in immediate vicinity	■	■	■	■	■	■	■	■	■
Age of interviewer	□	□	□	▒	■	■	■	■	■
Sex of interviewer	□	□	□	■	■	■	■	■	■
Weekday of first visit	■	■	■	■	■	■	■	■	■
Type of house respondent lives in	■	■	■	■	■	■	■	■	■
Respondent's sex	■	■	■	■	■	■	■	■	■
Number of people living regularly as members of the household	■	■	■	■	■	■	■	■	■
Children at home or not	■	■	■	■	■	■	■	■	□
Legal marital status	■	■	■	■	■	■	■	■	■
Years of full-time education completed	■	■	■	■	■	■	■	■	■
Main activity last 7 days	■	■	■	■	■	■	■	■	■
Household's total net income	■	■	■	■	■	■	■	■	■
Telephone number	■	■	■	■	■	■	■	■	■
Interviewer number of interviewer who started visits	■	■	■	■	■	■	■	■	■
Interviewer number of final interviewer	□	□	□	■	■	■	■	■	■
Calculated number of first contacts (workload)	■	■	■	■	■	■	■	■	■
Number of completed interviews	□	□	□	■	■	■	■	■	■
Workload divided by completed interviews	□	□	□	■	■	■	■	■	■

Table 3: Variable Availability. Dark grey: Variable Available in Respective Round Same Operationalisation Than in Other Rounds. White: Variable Not Available in Respective Round. Light Grey: Variable Available in Respective Round but Other Operationalisation, Format or Wording Than in Other Rounds.

3.5. Survey Weighting

The ESS provides design weights (DWEIGHT), post-stratification weights[8] (PSPWEIGHT) as well as population size weights (PWEIGHT) to correct for different inclusion probabilities due to the sampling design or differences in country populations. When analyses aim to generalise to the general population of a country based on the ESS data as a sample of this population or when comparisons between the countries' general populations based on ESS data are the aim of the analyses, then applying weights is mandatory. If weights are not applied, the sample does not represent the true marginal distribution of a target population if it is anything other than a simple random sample. Consequently, any generalisation made to the general population from such analyses would not be valid. A guide to ESS survey weights is available at the project website.

However, the following analyses do not aim to make inferences about the general population of a country, nor do they aim to compare the general populations of different countries. Instead, the raw samples are analysed and compared. Such an approach does not require the use of weights. In fact, weights are only available for participants of the survey but not for nonrespondents. Since nonrespondents are of crucial interest in this investigation, weights cannot be applied for this sub-population. Consequently, after consultation and in agreement with experts from the ESS Survey Weighting Panel, it was decided to not use any

[8] At the time of this analysis post-stratification weights were not available in version 1.2 yet (European Social Survey 2018b, p. 9).

weights in the following analyses. All results are referring to the sample of a country and not its general population.

3.6. Statistical Methods

Analyses in Chapter 4 and parts of the analyses in Chapter 6 focus on investigating the univariate distributions and bivariate relationships between the operationalised independent variables and the outcome variable. First, frequency and distribution analyses are presented for all variables and for all three countries. For categorical variables, the total number of observations per category as well as the category's share of the overall valid (i.e., non-missing) answers is reported. Arithmetic means as well as standard deviations are reported for continuous variables. Kernel density plots are created to give further insight into the distribution of the first contact hour. Heatmaps are used to visualise the relationship between the frequency of first contact attempts and the day of the week and time of the day.

To investigate potential relationships between independent variables and the outcome variable, bivariate analyses are reported in Chapter 4 and Chapter 6. For categorical variables *Chi*^{2}-tests (χ^2) will be presented, which tested for independence between a potential correlate of contact and the outcome variable. For simplicity, the perspective is adjusted slightly for continuous independent variables. Instead of running logistic regressions for each individual continuous variable and a combined model to interpret the influence of changes in the independent variables on the binary outcome, independent samples *t*-tests are applied to investigate group-mean differences of continuous independent variables between contacted and non-contacted units. Pearson's product-moment correlations (*r*) are calculated between all independent continuous variables. To also investigate the strengths of any relationships between categorical predictor variables, they are split into *k* dummy variables, where *k* is the number of levels of a categorical variable. Pearson's *phi*-coefficients (φ) are computed for the resulting 2x2 contingency tables. *φ*-coefficients for a 2x2 contingency table follow an intuitive interpretation like the one for Pearson's *r*-coefficients. Both Pearson's product-moment coefficients and *φ*-coefficients are presented in colour coded correlation matrices.

Results from χ^2 and *t*-tests were considered statistically significant at the 5%, 1% and 0.1% significance level. Correlations were considered statistically significant at the 5% significance level.

The specific machine learning methods for Chapter 5 and Chapter 6 will be outlined in Sections 3.8 and 3.9.

3.7. Software, Hardware & Open Code

Almost all data related operations for this thesis were deployed using R as well as RStudio Desktop for Ubuntu. LibreOffice Writer served as the primary word processor. Zotero has been used as a reference and citation management software and Linux Mint 19.3 Cinnamon as the operating system. All software was chosen deliberately to support and promote the use of Free and Open-Source Software (FOSS). All this software is free, which means that regardless of a researcher’s financial background, research can be conducted, and the source code is publicly available for inspection and change. Nevertheless, Microsoft Word was used for the final typesetting and editing of the thesis for convenience. Additionally, due to the complexity and required capacity load for the simulations in Chapter 7, the simulations were deployed on a different machine which runs R in Windows 10.

Crucial R packages included but are not limited to ‘haven’, ‘corrplot’, ‘ggplot2’, ‘sqldf’, ‘dplyr’, ‘caret’, ‘randomForest’, ‘e1071’, ‘glmnet’, ‘ROCR’ and ‘funModeling’. The Classification and Regression Training (‘caret’) package deserves honourable mention because without this package as well as its brilliant documentation (Kuhn 2019) the analyses would have been much harder to accomplish. A list of required packages can be found in the respective files in the GitLab repository.

To contextualise the computational time, which the algorithms in Chapter 5 and Chapter 6 needed to train the models, it is important to know the hardware specifications of the system the code was deployed on. All analyses were run on a MSI GL72 6QF-405 laptop with an Intel i7-6700HQ CPU (4x 2600MHz) and 8 Gigabyte of RAM. CPU benchmarks implemented via the R package ‘benchmarkme’ placed the hardware specifications in the midfield compared to all other available CPU benchmarks.

All code that is necessary to comprehend this analysis, i.e., for the creation of the datasets and plots as well as their analyses, is publicly available and can be inspected in the following GitLab repository: https://gitlab.com/wattseheck/dissertation.

3.8. Machine Learning Meets Survey Methodology Research

Machine Learning (ML) and Data Science are conceivably some of the most in vogue terms in the sphere of analytics today. Businesses seek to gain market advantages and large collaborations of researchers try to find the most elaborated mathematical models to make predictions better by the day. While interest and discussions involving machine learning, data science and their potential societal consequences have become more popular in data discourses in the last 10 to 15 years, the (mathematical) concepts behind these approaches are already well established. In fact, an Artificial Intelligence (AI) called 'Deep Thought' defeated master chess player David Levy already in 1989 (Hsu et al. 1990). To finally break through, ML only had to wait for technological advances to implement the concepts at scale.

The first big wave of research into machine learning took place between the 1960s and 1980s and some of today's most prominent algorithms like decision trees (Breiman 1984; Hyafil and Rivest 1976; Quinlan 1986), *k*-Nearest Neighbours (Cover and Hart 1967) and Support Vector Machines (SVMs) (Aizerman et al. 1964) were developed or have their roots in those years. Even the journal exclusively dedicated to machine learning (called *Machine Learning*) goes all the way back to 1986. The following two figures and their commentary might help to understand the rapid increase in interest. These are only two examples, but they convincingly reveal the importance of this topic today – both in and outside academia. Figure 1 shows the number of publications for the search string '(machine learning) *OR* (predictive modelling) *OR* (pattern recognition) *OR* (data science)' for the years 1970 to 2020 using the Web of Science database. While a minor increase in publications can already be observed in the 1980s, a massive and still ongoing increase in literature begins just after the year 1990. This increase correlates with major advances in computational capabilities, vast technological developments, and tremendous reductions in prices for computer hardware and storage. The first peak in Figure 1 in the early 90s correlates with both the birth of the 'World Wide Web' as well as with the development of Microsoft Windows 3.0 in the year 1990, which simplified the Graphical User Interface of computers. The second sharp increase in 1997 correlates with the shipment of Windows 95 in 1995 and Windows 98 in 1998[9]. The massive decline in publications from 2004 to 2012 is an outlier to this general trend and remains a curiosity. The fact that the cost of storing massive amounts of data has declined so sharply over the last decade, while computational power and advances increased at a staggering pace, might explain the almost vertical increase in publications in the years following 2012, since no one needed a 'supercomputer' anymore to apply machine learning but could instead use their everyday laptop. In 2019, there were a total of 79,810 publications in the Web of Science database, amounting to a 23% increase in publications compared to the year 2018.

[9] Also in 1997: Chess world champion Garry Kasparov was defeated by IBM's 'Deep Blue' chess supercomputer (Campbell et al. 2002).

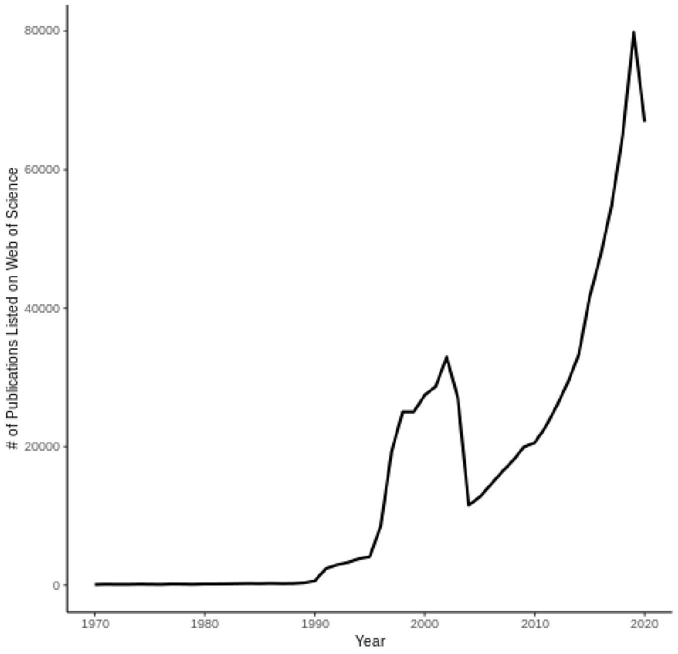

Figure 1: Number of Publications per Year for the Search String '(machine learning) OR (predictive modelling) OR (pattern recognition) OR (data science)' Listed in the Web of Science Database. Updated 26.11.2020.

Figure 2 shows the sharp increase in the popularity of the keywords 'machine learning' and 'data science' as Google search terms over time relative to their all-time popularity highs and lows between 1 April 2009 and 25 November 2020. The trendlines are illustrative of the speedy upsurge in the interest of these topics.

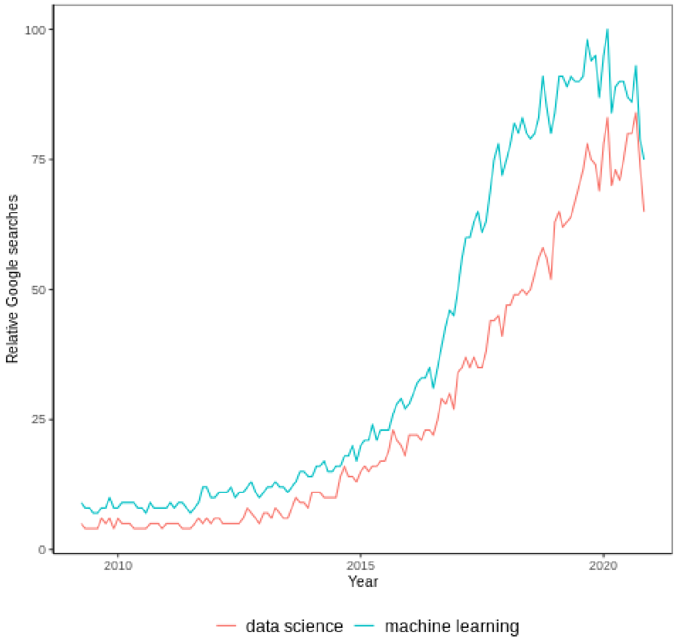

Figure 2: Proportion of Google Searches for Search Terms 'machine learning' and 'data science' Between 1 April 2009 and 25 November 2020. Updated 26.11.2020.

Despite these tremendous increases in research, publications and general interest in the topic, applications of methods that have their roots in the field of data science are still rare in the social sciences. There is, however, a growing community of scientists who promote the use of machine learning and other data science techniques (for example, deep learning or natural language processing) in social science research. It is difficult to estimate the size of this 'data science-savvy' group of researchers within the social sciences, since the applications are widely spread across a large amount of (quantitative) social science sub-disciplines. Consequently, using the number of publications would make for an overly complex and imprecise estimation. Nevertheless, there is no reason to assume that researchers in the social sciences are not interested in these developments and in applying cutting edge methods. This might especially be the case since the application of modern data science algorithms is increasingly facilitated by easy-to-use software by the day.

Although machine learning and data science techniques are getting more accessible, the roots of the field of data science, which lie in the field of computational sciences, usually bring their own set of different programming languages and analysis software and packages that are less common for typical social scientists. While quantitative social scientists are commonly using SPSS, Stata and R for their analyses, data scientists rely almost entirely on R or Python for programming. Although both SPSS and Stata can apply machine learning algorithms, frontiers in research are typically pushed in R or Python due to their open-source nature and large methodological and programming community.

Despite the differences between the fields, nowadays there are multiple dedicated machine learning, data science or artificial intelligence sessions at most large European or American social science conferences. Even a whole field of Computational Social Science with its own conferences, journals, university institutes (for example, the Chair for Computational Social Sciences at the RWTH Aachen University in Germany) and data science university tracks have emerged. Also, the field of survey methodology is an excellent example of such developments since it brings together multiple disciplines. When researchers from the substantive fields of sociology, sociology of knowledge, (social-) psychology, statistics and computer science collaborate, the survey methodology field profits a lot from these interdisciplinary synergies. Since survey methodology can be quite technical and mathematical (consider survey statistics, sampling, and weighting, for example) researchers in this field might have an affinity for applying new technical approaches to their field. At least for the last five years an emerge in publications and events which connect data science and survey methodology is visible. The books 'Big Data and Social Science' (Foster et al. 2017) as well as 'Big Data Meets Survey Science' (Hill et al. 2020) are only two examples of publications which contain comprehensive publications related to data science from numerous renowned authors from the field of Survey Methodology, which try to bridge the gap between the fields. There are multiple dedicated machine learning sessions at international social science conferences like the European Survey Research Association or the American Association for Public Opinion Research and even a distinct conference called 'BigSurv', dedicated to approaches of data science and big data in the social sciences, emerged.

Due to all these efforts to promote data science techniques such as machine learning, the growing interest in this topic and the technology at hand, it is no surprise that the field of survey methodology has already generated a fair body of literature applying data science techniques to survey methodology problems. While a large proportion of the literature focuses on predicting (non)response or participation in both cross-sectional and panel studies (Buskirk 2018; Eck 2018; Kern et al. 2019b; Kirchner and Signorino 2018; Kolenikov and Buskirk 2015; Liu 2020; McKay 2019; Signorino and Kirchner 2018; Würbach and Zinn 2019), other research focuses on substantial research questions like predicting voting behaviour (Bach et al. 2019), broad introductions e.g. to tree-based methods for survey researchers (Kern et al. 2019a) or more specific problems like automated coding of open-ended questions or question classification (Bullington et al. 2007; Chai 2019; Sangodiah et al. 2015; Zhang and Lee 2003).

Even though there is a large amount of research on participation and nonresponse, the process of contacting units is an under-researched topic in survey methodology studies that incorporate machine learning techniques. To the author’s knowledge, there are no studies to date that apply machine learning techniques to specifically predict first contact attempt success. Chapters 5 and 6 of this thesis try to fill this void. The purpose of this investigation is twofold. The first is of a technical nature, with efforts made to extend the foundational research by contributing to the discussion of whether and how machine learning techniques are useful for answering survey research questions. Second, the investigation aims at extending the substantial insights of the contact process and investigating whether predicting first contact attempt success is feasible in the context of the ESS. This investigation lays the foundation for the prototype of a simulation approach, outlined in Chapter 7, to tailor the attributes of a first contact attempt to the needs and traits of a target unit. This prototype could be extended by future practitioners and researchers to develop survey fieldwork tools that may improve fieldwork efficiency. One application might be to use an advancement of the prototype in conjunction with the Fieldwork Monitoring System, which was used in the ESS Round 9, or as the decision-making algorithm inside a micro-simulation of fieldwork processes to decide which units in the simulation are successfully contacted at first attempt and which are not. The concept of such a micro-simulation was already presented in detail by Schnell in 1997, who pleaded for an approach that enables researchers to vary fieldwork conditions in a laboratory environment (Schnell 1997, p. 242-244). Although Schnell points out that that the concept is not meant to be a prediction, it might be worth investigating whether the concept can be amplified with the help of predictive modelling (Schnell 1990).

Thus, the investigation is aimed at both technical survey methodologists with an interest in data science as well as practitioners and fieldwork managers who are trying to increase the efficiency of their fieldwork. The development of a successful prototype could prove as a starting point for other researchers, who could enhance it and validate its feasibility by applying it in a real-world survey fieldwork environment. The prototype, which carries the potential to significantly reduce the costs of initial contact procedures or at least estimating them more precisely, can then be adapted, applied and extended to each specific survey environment. Reductions in overall survey costs would make conducting surveys more economically efficient and enable fieldwork agencies to allocate resources to other crucial components of survey fieldwork that are likely to increase data quality – for example, raising interviewer salaries. The prototype is driven by successful machine learning algorithms, which will be introduced alongside other important concepts in the next sections.

3.9. Demystifying Data Science for Social Scientists

Some of the methods and terminology used in Chapters 5 and 6 will appear to be different from those which are commonly applied in the social sciences. To demystify some of these concepts, the following sections are explicitly dedicated to show that there are obvious parallels between techniques commonly used in the social sciences and data science. By giving a gentle introduction to readers with a background in the social sciences, it will be pointed out that the biggest difference between the fields of quantitative analysis in social sciences and machine learning in data science is predominantly a matter of perspective and scope of research.

The aim of Chapter 5 and 6 is to find an approach of *predicting* whether a previously unobserved unit can be successfully contacted at first contact attempt in the future based on model specifications which were derived from historic data. To achieve this, the analyses will make use of machine learning methods which leverage data from the ESS.

At a first glance, the differences in terminology, methods or software used between social science and data science can seem large. This may be discouraging for researchers, practitioners or students and may prevent them from exposing themselves to unfamiliar approaches and incorporating them into their research or day-to-day work. Yet, by taking a closer look, one not only finds out that the concepts are more similar and applied much more easily than anticipated, but also that a lot of useful approaches can be adopted from a machine learning workflow in addition to traditional statistical analyses.

To give a brief overview of what the next sections will elaborate on in more detail: in a first step data is split into a training and testset. Then machine learning algorithms are used to estimate an internal score (often a probability) for each unit based on the characteristics of their input variables before they assign each unit to one of the two possible outcome classes: ‘no contact at first attempt’ or ‘contact at first attempt’. In most algorithms, the assignment of a unit to either of the outcome classes happens in accordance with a user-specific or algorithm-specific threshold of their probability. This might remind social scientists of logistic regression. In fact, similar principles apply. In a logistic regression model, the probability of belonging to a group (0 or 1) is estimated. The algorithm-specific threshold to belong to either of the groups in a logistic regression is usually fixed to a probability of 0.5, meaning that observations with an estimated probability smaller than 0.5 are classified as 0 while those with a value greater than or equal to 0.5 are classified as 1 (Best and Wolf 2010). After the machine learning algorithm assigns each unit to one of the groups, the algorithm’s predictive performance needs to be evaluated, i.e., whether the predictions were actually correct. There are plenty of different performance indicators, but we will establish the concepts of *sensitivity* and *specificity* as well as the prediction's specific so-called *Receiver-Operator-Curve* in a later section.

The kind of analyses and terminology in Chapters 5 and 6 might be unfamiliar to those researchers with a background in the social sciences. Because of this, the next section first gives a practical and gentle introduction to the methods and terminology of machine learning and data science. It is meant to be presented and explained in a way to be a helpful guide for social scientists who encounter machine learning or data science for the first time or have little knowledge about them. Before any analyses are run or any predictions are made, it is the aim of this section to equip readers with the relevant knowledge that is needed to follow along and understand the later chapters of this thesis but also data science more generally. However, this section is not meant to replace any comprehensive literature of machine learning and data science.

As this section evolves, the focus zooms in from a high-level perspective of comparing the fields of data science and social science to digging deeper with each chapter into what happens inside the machine learning workflow. Firstly, it is important to get a clearer picture of what machine learning tries to accomplish and how this might differ from a social science analysis. Then, some crucial terminology is introduced, transitioning from floating over the topic to diving into what actually happens inside the data training process. After investigating resampling techniques and hyperparameter tuning in more detail, we look at how performance is evaluated. At that stage, we have seen all relevant steps in a machine learning workflow and the chapter closes with an introduction to five algorithms that are used in the later analyses.

3.9.1. A Matter of Perspective

Let us start by comparing the fields of data science and social sciences from a macro-perspective. It can be argued that one difference between the fields lies in the perspective from which a problem is looked at or what the analysis tries to accomplish i.e., the scope of the analysis. In most quantitative social sciences, it is the researchers' primary objective to apply analyses to *explain* relationships between independent variables (the *predictors* in machine learning terminology) and a dependent variable (the *target*). A quantitative social science analysis can thus be used to describe how the outcome changes as a function of its inputs based on previously collected data on units and a carefully specified model. In traditional statistical modelling, assumptions on the data are usually made a priori and the modelling can be influenced by the researcher, who makes certain modelling decisions based on previous research or literature (for example preferring particular interaction terms over others). (Social) phenomena can then be explained by making use of inferential statistics (and a previously applied sophisticated sampling strategy) and generalising from the sample results to an inferential population. Additionally, these analyses can also be used to determine how strong a specific component of the model influenced the dependent variable and in which direction. To illustrate, in an analysis of party preference, a typical research question could be to try to explain how sex, union membership status or age as well as an interaction effect between sex and union membership influenced a specific voting intention and which of these independent variables had the largest impact on an election in retrospect.

In contrast, machine learning tools are usually not used to *explain* a relationship i.e., to find out which predictor influences the outcome, to what extent and in which direction, or to infer from a sample to make statements about an inferential population. Instead, machine learning algorithms are used to model the underlying data structure by re-iterating over the data over and over again, finding specific cut-points and thresholds themselves and thus learn from the data itself to find the best model that *predicts* the future outcomes of unobserved data points. Therefore, little to none input is needed from the researcher[10]. Usually there is less focus on investigating specific model parameters than in traditional statistical modelling. For example, a specific interest group might be interested in how many Labour members the future parliament will have. As a first step this interest group gathers data on all past elections. Since it is more important to know *that* the future parliament will have *x*-Labour members, based on a prediction of voting intentions, than knowing *why* people voted the Labour instead of the Conservative Party, they apply a predictive modelling algorithm, which models the underlying data structure and searches for specifications that are able to predict the outcome of the future election with lessor no focus on explaining the voting intentions.

Just like in traditional statistics this process can become complex and difficult to understand what happens inside the algorithms. Additionally, since the primary focus lies on prediction, it is often not practically useful (or comprehensible) to derive any explanations from the algorithm like in traditional statistical analyses as the algorithm optimises predictive power

[10] It is often argued that machine learning models are less susceptible to biases from the researcher. While it might be true that less manual model building is involved, the argument for a lesser risk of biases remains highly controversial.

and not interpretability. Despite recent advances (Arras et al. 2017), for most complex algorithms, like Artificial Neural Networks, the technical and mathematical procedures, for example, for correctly weighting the different input variables in each so-called *layer*, become so complex, that an interpretation of what actually happens inside the algorithm is often considered a black box (McGovern et al. 2019; Zhang et al. 2018). If these concepts are explained in a textbook, they mostly refer to rather low dimensional data and examples which are far away from the complexity of real-world applications. These complex algorithms excel both in regression and classification tasks, and for example, in image recognition. However, explaining what exactly happens in the model is greatly demanding and reverse engineering the relationships between the predictors and the target, like in a social science research question, is presumably time-consuming and demanding (Arras et al. 2017).

At second glance, machine learning approaches and traditional statistical methods are deeply connected and separating them becomes more difficult. Still, it might be argued that they serve different purposes. This also implies that neither is superior to or more useful than the other. However, when it comes to making data-driven statements about the outcome of an unobserved data point, machine learning approaches are paramount since their aim is specifically to generate models that can do precisely this one and only job. They are not designed to be used as tools to retrospectively investigate relationships in datasets. Consequently, machine learning approaches should not be seen as rivals to methods from traditional statistics but rather a useful addition.

It is important to acknowledge these differences and similarities, and that researchers from the two fields might have a different perspective on their problems or research questions.

Personally, when I first exposed myself as a survey methodologist to learning about data science, the two fields appeared so different and the variations in the scope of an analysis appeared so large, that combining the two fields was challenging. Especially the differences in terminology hampered my self-learning process and understanding in the beginning. It was only after I found out that the fundamentals are similar or at least stem from the same roots, that I realised that the real differences were mostly in the terminology of the fields. After spending some time learning more about data science, I was able to see parallels between the fields but also realised that the touchpoints between the disciplines (at least in my personal scientific bubble) were just not large enough yet. I was and still am sure that collaborations between the fields are mutually valuable for both. In this thesis I want to help bridging the gap between them.

3.9.2. Data Science Terminology

The term 'machine learning' itself can serve as a starting point to dive deeper into some of the data science terminology. The word 'machine' hints at something that is done automatically by technology i.e., a computer. Automation does not imply that everything happens without the need of someone to supervise the processes at all. Usually, a researcher needs to 'tell' the machine what exactly needs to be automated. Further, the word 'learning' refers to an iterative process in which an algorithm tries to find the best possible solution for a given specific outcome criterion. In this learning process, several combinations of possible solutions are compared against the outcome criterion and the most successful is chosen as the optimal solution.

Machine learning algorithms extend traditional analyses by the ability to *predict* outcomes for unobserved data points as correctly as possible. This prediction is a logical consequence of solving complex mathematical equations and is ultimately the estimated outcome of an equation for units whose outcomes *were not* observed based on the estimates of an outcome for units whose outcomes *were* observed. Strictly speaking, this only applies to a certain kind of machine learning: the so-called *supervised learning*, as detailed further below. Just like in traditional statistics, one can distinguish two types of machine learning by looking at the outcome variable, which is often referred to as the 'target' variable. These two types of machine learning are sometimes referred to as the two different kinds of 'machine learning problems' that need to be solved. Identifying and distinguishing between these two problems is facile for social scientists since they are no different from traditional social science methods: when the outcome of the prediction – the target variable – is a categorical variable, whether binary or multi-class, a so-called *classification problem* needs to be solved. On the other hand, when dealing with a continuous outcome, methods for *regression problems* need to be applied. This distinction is familiar to statisticians, who would apply logistic regression models for categorical outcomes or linear regressions for continuous measures, respectively. The only challenge might arise from the word 'regression' in 'logistic regression' which is not related to solve a regression problem in machine learning but instead a classification problem. Like in traditional statistics, there are dedicated methods to solve either classification or regression problems, but there are also methods that can be applied to both. This example is the first of many that highlights that the differences does not lie in the method but in the terminology between the fields.

Coming back to the term 'supervised learning', besides the two types of problems, there are also two different approaches in machine learning with their own very specific rationale: *supervised learning* and *unsupervised learning*. First, in supervised learning, the so-called *class label* is available in the dataset. This class label is a string or value of a particular outcome variable of interest. In a classification problem context, this variable contains the information in which of *n* classes an observation belongs whereas in a regression context this variable contains the information of some continuous variable. For example, when trying to predict party preference, researchers make use of a survey dataset, which features different variables on political questions – including the party preference of the surveyed units – and demographics. In addition to predicting the party preference of surveyed units, one is also able to predict the party preference for units that were not observed in the survey based on the available information from the observed units in the dataset. From the survey, one might have data on the observed units' sex, age, their union membership status and their party preference as the class label. A model of choice can be *trained*, which uses these variables (sex, age, and union membership status) as predictors for the target 'party preference'. For each unit in the dataset, the model will solve a classification problem, predicting a value and assigning each unit to exactly one category within the list of possible parties. Thus, such modelling gives each unit a predicted class label. The enhanced dataset now contains the true observed party preference as well as the predicted party preference. The success of the prediction can now easily be measured by validating the prediction against the actual party preference. By looking at a crosstabulation between the predicted class-label and the observed class-label of the party preference, the success of the prediction can be evaluated on a wide range of metrics, which will be explained later. One could say that the aim of supervised machine learning is to find the one model that is able to predict the information of a target variable that is observed already. By comparing how successful this prediction was with the actual observed value, the algorithm can train itself iteratively to find better solutions. Once the algorithm finds the best solution for the observed data, it can then be estimated how well the algorithm will perform when it is applied on new data, where the same predictor variables were surveyed, but the target variable was not observed.

Second, in unsupervised learning, approaches usually try to bring order and structure to the vast amount of data a researcher might deal with. Here, the class label of a target variable is not present in the dataset and a prediction cannot be validated using an observed class label. In fact, one does not even necessarily know which observations belong together, thus unsupervised machine learning is used to group and structure these data points based on their characteristics of interest. The idea of the concept might be familiar to social scientists due to their exploratory character, which is also useful and common in social science projects. In fact, different variations of cluster analyses are common unsupervised learning methods. The primary aim of unsupervised machine learning techniques is to reduce the complexity of large datasets. For example, imagine a virtual folder with numerous articles that necessitates sorting these documents by their topic and storing similar documents together. Usually, papers do not come with one single class label, which reveals the one and only topic they are about, so predicting the class label or a quantitative value and comparing this prediction to the actual value is not an option. Instead, models that search for similarities and differences between the documents can be applied to group them. These algorithms might, for example, compare the documents based on the type and number of words that appear in them. In such an approach, all the papers would be transformed into a tidy dataset, with one row containing the paper that is referred to identified by an index, columns indicating whether a specific word is mentioned or not and unnecessary fill words deleted. Afterwards the algorithm would examine the dataset for similarities in the word occurrences. This might, for example, be done by applying distance measures on the dataset such as the *k*-Nearest-Neighbours (KNN) algorithm, by assigning more similar documents to the same group. This distance measure may be familiar to social scientists in the context of Principal Component Analysis (PCA). Since, in contrast to supervised learning, there is no class label, which can be used to validate whether this predicted grouping was correct, one option is to check for internal and external consistencies. One important criterion might be that there is high homogeneity within the groups and high heterogeneity between the groups, to make sure the grouping or clustering worked as intended.

This introduction to some of the basic terminology should suffice to dive deeper and explain how machine learning models are trained in the next section.

3.9.3. Training and Resampling a Model

Machine learning can be divided into classification and regression methods as well as into unsupervised and supervised techniques to classify or to predict an outcome rather than to explain relationships. Most of the following remarks focus on examples of supervised machine learning since it is of more relevance for this research. A crucial phase in machine learning is *training a model*. In traditional statistics, researchers commonly refer to *fitting* a model e.g., to describe a relationship in the context of regression modelling. In machine learning, the equivalent term to *fitting* a model is *training* a model.

Since data scientists often work with a vast amount of data at hand, an approach to handling that data would be to train an algorithm on the whole set of available data, as is usually done in traditional statistics. However, if an algorithm was only trained using the entire set of data, without validating the algorithm on a second one, it would be susceptible to *overfitting.* Overfitting means that the estimated function to predict the outcome follows the observed data too closely and not only models the underlying structure, but also predicts the errors (noise or bias) in the data. This means that an overfitted algorithm that follows patterns of the trainset too closely would perform satisfactorily in the trainset but poorly on the testset. Splitting the available data leverages this trade-off. To evaluate an algorithm's performance and to prevent overfitting, an algorithm is not trained on the whole available data but only using a subset of the available dataset. In other words, the full dataset is split, and the algorithm is trained using only one subset (*the training set)* and tested in another one (*the testset*). The approach here is to compare the prediction for an observation in the training set against its observed outcome. After the algorithm has found the best specifications for the training data, these specifications are applied on the testset. Since the data from the testset was not used to train the model, it can be used to evaluate how well the algorithm performs when applied to new but identically structured data.

There is no strict threshold to determine the size of the two subsets. A common approach in machine learning is to train the model on 70 percent of the original dataset and test it on the remaining 30 percent, but 75/25 or 80/20 splits are not uncommon. More important than the actual size of the partition is that the training set includes substantially more data points than the testset since the modelling is performed using the former while the latter is 'only' used for evaluation purposes. Furthermore, it is important that the split does not introduce any biases in the two sets but instead each observation is assigned randomly to either the training or test partitions.

The next paragraphs focus on what happens inside the training set. The so-called *resampling* methods are a set of techniques – slightly familiar to traditional social scientists but used in a different way in machine learning – that come into play only in the training set. Resampling is done by splitting the training set into even smaller subsets. Instead of training the model on the entire training dataset at once, the training data itself is split into two or several smaller subsets for the algorithm to be trained on each one of them. Put simply, the training data can be easily seen as internally split – or resampled – into a train-trainset and a train-testset. The idea is for the algorithm to be trained on one of the subsets (the train-trainset) while the other(s) mimic a testset (train-testset) within the trainset before the next iteration begins and another subset becomes the trainset. With the help of this approach, it can be observed how the estimated so-called *test error* – as explained further below – differs from subsample to subsample and the algorithm draws conclusions from them to find the best model. Afterwards the gained information and estimations on how the algorithm would perform on new data are averaged and the best performing hyperparameters are fed back.

The correct term for the train-testset is the *validation set* or *hold-out set,* which also gives this resampling procedure the name *validation set approach*. In its simplest execution, this approach splits the trainset randomly into two halves. The model is trained using the one half and a validation set error is calculated for the other half – the validation or hold-out set. The validation set error then serves as an estimator for the test error of the final testset. There are two major drawbacks of the validation set approach. The first lies in the high variability depending on which units are randomly chosen to be in the train-trainset and the validation or hold-out set. Second, in a 70/30 split for the train and test data, the trainset already has a reduced sample size of only 70% of the original dataset. If it gets further split into two equal halves, the sample size is reduced even further and the model is trained on only 35% of the original sample, which increases the likelihood of overestimating the test error rate based on the validation set error (James et al. 2013, p. 178).

Thus, to overcome these issues *cross-validation* is a more enhanced approach that has evolved from the validation set approach and which tries to overcome the issues of its predecessor. Cross-validation can be applied in multiple ways. One approach is the so-called *Leave One Out Cross-Validation* (LOOCV) or Jack-knifing, which does exactly as the name suggests: it leaves out one observation for every iteration. In a trainset of size *n* the model is trained using a train-trainset of *n-1* observations and the remaining one observation is used as the validation set. The strength of this approach is the fairly unbiased estimation of the test error since almost all observations are used to train the model. On the other hand, fitting models *n-1* times can be computationally expensive and inefficient.

To overcome the downsides of LOOCV, the so-called *k-fold cross-validation* approach was developed. In the light of k-fold cross-validation, LOCCV can be considered a special or extreme case, where *k* is set equal to *n-1*. In k*-fold cross-validation* (KFCV) the train data is split into *k* different subsets or folds – usually *k=5* or *k=10* – of roughly equal size. Each one of these folds is treated as the validation set for one training iteration before the next fold is used as the validation set and the model is trained again. Instead of repeating the process *n-1* times like in LOCCV, only *k* repetitions are needed. This approach then leads to *k* different validation error estimates that are averaged to approximate the test error rate. The model that produces the lowest estimate for the test error, is considered to be the best performing model and can be validated against the test data. KFCV combines the strengths of LOOCV (being more unbiased than the validation set approach and having no random splits of data) and adds better computational performance since not *n-1* models but only *k*-folds need to be computed, which increases efficiency. Furthermore, KFCV is considered to have some statistical advantages over the LOOCV: while LOOCV has a lower bias than KFCV, since almost all training observations are used to train the model, it is susceptible to a higher variance than KFCV. This is because the validation set error in LOOCV is built by averaging *n-1* validation set errors, which are based on almost identical trainsets. These errors are considered to be highly correlated with each other and consequently higher correlated than those from KFCV, leading to a higher variance of the LOOCV test error rate (James et al. 2013, p. 183f). Hence, 5-fold and 10-fold cross-validations not only have computational advantages but also yield more accurate estimates for the test error rate.

To further improve the advantages, this *k*-fold cross-validation can be repeated: in a *repeated k-fold cross-validation,* the trainset is not only divided into *k*-folds once, but instead is divided into *k*-folds repeatedly for *x* – typically 5 or 10 – times. At each of these repetitions, a different starting observation is chosen to create different subsamples of the dataset and to minimise the risk of any biases due to any unobserved underlying structure in the dataset. For this reason, all the analyses of the next chapters will use repeated k-fold cross-validation as a resampling technique.

By applying resampling techniques, the trainset gets internally split even further to be able to estimate the test error rate explained later in this section. The hyperparameter combinations for an algorithm that produce the lowest predicted test error rate are considered the best hyperparameter combination and are chosen as the 'winner' to be validated against the test data. The process of finding the optimal combination of hyperparameters for an algorithm is called *hyperparameter tuning*. Hyperparameter tuning is the process that is hidden behind the word 'learning' in the term 'machine learning' and will be explained in the next section.

3.9.4. Hyperparameter Tuning

Similarly to models like linear regression, most machine learning algorithms possess coefficients that can be altered to find the most optimal solution. The machine learning terminology calls these coefficients *hyperparameters*. Algorithms can have different numbers of hyperparameters that need to be tuned to make the algorithm work. A full list of algorithms, which are supported by the caret package as well as the corresponding tuning parameters for each model can be found in the caret documentation (Kuhn 2019). The range of these hyperparameters is set *before* a model is trained using resampling techniques in the training process to find the best solution. Setting the range of hyperparameters before the training procedure is important because they are the one central leverage which defines the training process of a model. Speaking of 'a' hyperparameter can be misleading. Most machine learning models have several hyperparameters and each of them is not determined to one specific constant value *v*, but instead the algorithms are fed with different value ranges or vectors of values $\begin{pmatrix} v_1 \\ \cdots \\ v_m \end{pmatrix}$ for each hyperparameter. The algorithm can cycle through this vector of possible hyperparameter values in the training process and test out multiple different combinations. By trying out these different value combinations of hyperparameters, the machine learning algorithm is looking for the single best combination to estimate the test error rate. The range of vector values can either be set from a manual and predefined list of values or fed to the algorithm in a random process.

Imagine a cloud of data points of two classes in a two-dimensional sphere that are linearly separable by a function of $f(x) = b + mx$. If one desires to find the best possible linear separator – the one that maximises the distance between the classes to reduce the risk of misclassification – different values of b and m as well as their combinations can be tried out to create different linear functions with different slopes and intercepts in the two-dimensional sphere. These different combinations could be compared and evaluated against each other to find the best separator manually. To avoid this manual tuning, it is possible to feed modern software and packages like caret with so-called *tuning-grids*. These *tuning-grids* contain vectors of different values for both *b* and *m*. The algorithm then uses only the values from this predefined tuning-grid and feeds them to the corresponding hyperparameter *b* or *m*, respectively. The benefit of this approach is the complete controllability of the tested hyperparameters by the researcher. Afterwards, it applies the resampling and trains the model using the trainset and evaluates which combination of these values leads to the optimal outcome of a predefined evaluation metric, e.g., the Receiver-Operator-Curve (ROC), explained later, in the train data, which is predicted to also lead to the best outcome in the testset.

Besides the possibilities of manually trying out hyperparameters or using tuning-grids, there are also ways of providing vectors of values to each hyperparameter of an algorithm, which work more autonomously but lack controllability. Both a manual selection and using a tuning-grid of predefined values for hyperparameters can be considered sub-optimal. Since the researcher needs to explicitly define a satisfying parameter selection from an infinite number of possible hyperparameter combinations, the chosen combination might potentially ignore the most optimal specification by chance, due to a lack of knowledge or due to a selection bias if a researcher just chooses the numbers manually. One approach to dealing with this downside is to let the software search for random values and test combinations of these random values instead. In theory, the random hyperparameter search finds the best specification when the search goes on long enough. While random hyperparameter search deals with the problem of subjectivity in the process of manually setting the hyperparameters, it also comes with two obvious downsides itself. A random hyperparameter search not only is costly in terms of computational time, but also it is possible that a random search does *not* select the best values by chance if the random search does not go on indefinitely but is stopped before the best hyperparameter was found.

To combine the strengths of computational efficiency and a random search, *adaptive resampling* has been developed. Instead of looking at hundreds or thousands of random values, adaptive resampling takes a starting value, then creates a second value and tests whether the second value combination performs better or worse. This depends on which of them produced better training-set results. Similar to the maximum likelihood estimation, it then looks for new values in the direct neighbourhood of the better performing values and makes a new comparison. The adaptive approach is equipped with a confidence level indicator that is commonly set to 5%, which means that the algorithm must be confident on a 5% error-level to not discard a hyperparameter combination. Adaptive resampling approximates the best hyperparameter combinations more efficiently. Although this approach has major strengths, one important downside might be that the algorithms can settle down on a local rather than a global optimum by chance, depending on the random starting point of the hyperparameter search. Yet, this approach is able to speed up the computation process tremendously, while having a higher chance of finding the most optimal tuning parameter value. For these reasons, the later analyses will make use of this method for hyperparameter search.

To sum up the description of hyperparameters, one could say that they are components of a recipe for a function or algorithm that need to be pre-defined in order to instruct the algorithm how to run. After the algorithm knows from which value space it can choose, it iterates through all these combinations and feeds back those combinations that yield the highest predictive power given a certain outcome criterion. A more technical read on adaptive resampling can be found in Kuhn (2014).

3.9.5. Evaluating Performance

This section will introduce the reader to the concepts of how the performance of an algorithm is evaluated. As explained in Section 3.9.3, the first step of the process is to train the model using the trainset to avoid overfitting. The section highlights that the trainset is further split internally and the results from the training are compared against a validation or hold-out set inside the train data to evaluate the estimated test error rate, before the best hyperparameter combination is fed back to be tested on the testset. Following this process, there are steps in which evaluation become important: one within the process of training the trainset and one to evaluate the performance of the best algorithm using the testset. Both evaluations are important. However, it is usually of less interest to know the exact training performance as long it provides the best estimate for the test data. Considering the voting intention example, from a machine learning perspective, it is of less interest to precisely predict whether someone voted Labour or not in the previous election (high performance in train data), since data from this election is already historic, than to know what a person sharing similar characteristics will vote for in the next election (high performance in test data).

Starting with the evaluation of the training set performance, the algorithm needs to decide which specifications are producing the best predictions for the testset based on the training data. The hyperparameters associated with this best prediction are fed back to be applied on the test data. The training data $(x_{1j}, v_{1j}, y_{1j}), (x_{2j}, v_{2j}, y_{2j}), \ldots, (x_{nj}, v_{nj}, y_{nj})$ with some predictor variables *x* and *v* and a target variable *y* for *n* cases in the trainset *j* are trained to predict the outcome $\widehat{y_{nj}}$. For each training observation, the individual target $\widehat{y_{1j}}(x_{1j}, v_{1j}), \widehat{y_{2j}}(x_{2j}, v_{2j}), \ldots, \widehat{y_{nj}}(x_{nj}, v_{nj})$ can be predicted. If this predicted outcome is close to the unit's observed outcome $y_1, y_2, \ldots, y_n$, then the training performance indicator – for example, the Mean-Squared Error (MSE) – is small, which hints at a good training performance. However, in the end, a researcher is less interested in the performance of the *trainset* or whether $\widehat{y_{nj}}(x_{nj}, v_{nj}) \approx y_{nj}$ in the trainset, but instead wants to know whether $\widehat{y_{0\imath}}(x_{0i}, v_{0i})$ is approximately equal to y_0, where (x_{0i}, v_{0i}, y_{0i}) are the predictor and target variables of a new observation, 0, from the *testset*, *i*, that has not been used to train the model. Consequently, the best model minimises the MSE for the *testset* or

$$min\big(mean((\widehat{y_{0\imath}}(x_{0i}, v_{0i}) - y_{0i})^2 + \cdots + (\widehat{y_{n\imath}}(x_{ni}, v_{ni}) - y_{ni})^2)\big) \tag{5}$$

for *n* observations in the testset (James et al. 2013, p. 30).

From the previous section one might think that partitioning the data into train and test data solves the problem of overfitting, i.e., modelling the data structure too closely to the trainset. This is not strictly true. However, overfitting can at least be evaluated if some sort of dataset partitioning is conducted, whereas the potential overfit remains completely undiscovered if the full dataset is used for training purposes. An overfit can be detected by comparing the MSEs for both the train and testsets: if an algorithm yields a small training MSE but a large test MSE, the algorithm is likely overfitted. That means the algorithm excels in modelling the data of the training set, but the resulting hyperparameters indicate an inefficient or biased prediction when this algorithm is applied to model new test data.

Taking a closer look at the MSE, for a specific observation, the MSE can be decomposed into the sum of the variance and squared bias of a variable – similarly to Equation 1:

$$\left(y_{0i} - \hat{y}(x_{0i}, v_{0i})\right)^2 = Var\left(\hat{y}(x_{0i}, v_{0i})\right) + \left(Bias\left(\hat{y}(x_{0i}, v_{0i})\right)\right)^2 + Var(\varepsilon) \qquad (6)$$

and can be generalised to the average MSE by averaging $E(y_{ni} - \hat{y}(x_{ni}, v_{ni}))^2$ over all n observations of the testset i. A model that tries to minimise the MSE, regardless of whether in the train or test MSE, needs to reduce both $Var(\hat{y}(x_{0i}, v_{0i}))$ as well as $(Bias(\hat{y}(x_{0i}, v_{0i})))^2$ terms but can never be lower than $Var(\varepsilon)$, which is the irreducible variance of the error term. A good performance can, thus, only be achieved if both variance and bias are reduced as much as possible. Unfortunately, there is a trade-off between variance and bias. More flexible algorithms, like Support Vector Machines (SVM), approximate a real distribution better and reduce bias. Unfortunately, they yield increased variances and are harder to interpret. On the other hand, less flexible algorithms (like random forests) result in higher bias as they cannot approximate a distribution as well but in return, they are easier to interpret and have less variance. As a result, the best applied machine learning algorithm tries to find the best balance between variance and bias to reduce the MSE in the testset (James et al. 2013, p. 34f), so this yields a solution that is as precise and accurate as possible, respectively.

The MSE is used as a performance indicator for regression problems, but similar concepts exist for classification problems. Instead of looking at the MSE for a numerical outcome, the *average training error rate* is the metric of choice in classification problems for a categorical or binary outcome in the training data. The *average training error rate* is the fraction of incorrect classifications defined as $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y_i})$ with $\hat{y_i}$ as the predicted class label, for the ith observation, and $I(y_i \neq \hat{y_i})$ as an indicator variable that turns 1 if $y_i \neq \hat{y_i}$ (incorrect prediction) and 0 if $y_i = \hat{y_i}$ (correct prediction). To evaluate the testset performance instead of the training set performance, the algorithm tries to find the minimum average of incorrect classification for the testset or $min(mean(I(y_0 \neq \widehat{y_0})))$ (James et al. 2013, p. 37f). The error rate is only one of several other important metrics that measure the performance of classification predictions and can be summarised in a *confusion matrix*. In its simplest form i.e., in a binary classification, the confusion matrix is reflected as a 2x2 cross-tabulation and represents the predicted and observed values as shown below.

Observed (y_j)	**Predicted** ($\hat{y_i}$)		**Total**
	Yes	*No*	
Yes	218 (TP)	57 (FN)	275
No	28 (FP)	179 (TN)	207
Total	246	236	482

Table 4: Example Confusion Matrix.

This example includes 482 observations in total. The observed class y_j (*yes* or *no*) can be found in rows whereas the predicted class $\hat{y_i}$ (*yes* or *no*) can be found in columns. This table identifies those observations which were predicted to be positive (*yes*) and were indeed positive (*yes*) in the top-left, light-green corner. Predictions that are correctly classified as the positive group are called *true positives* (TP) since their prediction to be positive is actually true. Diametrically opposite are those observations which were predicted to be negative (*no*), which were indeed observed as negative (*no*), in the bottom-right dark-green corner. Predictions that are correctly classified as the negative group are called *true negatives* (TN) since their prediction to be negative is actually true. The remaining dimension is defined by two cells containing those predictions that were incorrect. The bottom-left orange corner represents the so-called *false positives* (FP). *False positives* are those observations which were predicted to be positive (*yes*), while in fact they were observed as negatives (*no*). Contrary, the top-right red corner represents the *false negatives* (FN). *False negatives* are those observations which were predicted to be negative (*no*), while in fact they were observed as positive (*yes*).

These concepts are not new to traditional statisticians. False positives are also known as the *Type I error*. If an experiment tries to predict whether someone will answer a specific sensitive question (positive outcome) or not (negative outcome), and the test predicts the person to answer positively, while in fact they do not, a *type I error* is made, since the test falsely predicts the positive outcome for this observation. On the other hand, false negatives are known as the *Type II error*. A t*ype II error* is made when someone who actually answers the question positively gets a negative prediction. To provide a medical example: if a woman in her eighth month of pregnancy is told that she is not pregnant, likely a type II error was made. If instead a biologically male person is told that he is pregnant, this is likely a type I error (Trochim 2005, p. 207f).

The error rate and many other important rates can be derived from the confusion matrix. The error rate is defined as the proportion of classifications that were incorrect. Making use of the hypothetical data from above, the error or misclassification rate can be calculated as $\frac{(FN+FP)}{Total}$ or $\frac{(28+57)}{482} = 0.2$. Hence, our prediction is incorrect in 20% of our classifications. Some other important statistics can be derived from the confusion matrix (Fawcett 2006). These, amongst others, are:

- Sensitivity: $\frac{TP}{TP+FN}$. Also called *recall*, *hit rate* or *true positive rate*, shows the fraction of true positive predictions relative to all positive observed outcomes.
- Specificity: $\frac{TN}{TN+FP}$. Also called *selectivity* or *true negative rate*, shows the fraction of true negative predictions relative to all negative observed outcomes.
- False positive rate: $\frac{FP}{TN+FP}$, shows the fraction of false positive predictions relative to all negative observed outcomes.
- Precision: $\frac{TP}{TP+FP}$. Also called *positive predictive value*, shows the fraction of true positive predictions relative to the total number of all positive outcomes and measures whether the algorithm is accurate when it predicts the positive class.
- Negative predictive value: $\frac{TN}{TN+FN}$, shows the fraction of true negative predictions relative to the total number of all negative outcomes.
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$, shows the fraction of correctly predicted cases relative to the total number of observations that needed to be classified; in other words, how often the classifier is right.
- Prevalence: $\frac{FN+TP}{n}$, shows the fraction of the positive class's occurrence relative to the sample size.

Using the above example, the following metrics can be calculated and rounded to: *sensitivity* = 0.8, *specificity* = 0.7, *false positive rate* = 0.2, *precision* = 0.8, *negative predictive value* = 0.9, *accuracy* = 0.8 and *prevalence* = 0.5. The fictional prediction predicts 'positive' when the outcome is actually positive in 80% of the cases (*sensitivity*), 'negative' when it is actually negative in 70% of the cases (*specificity*) and 'positive' when it is actually negative in 20% of the cases (*false positive rate*). In cases when it predicts 'positive', this prediction is actually correct in 70% of the cases (*precision*), when it predicts 'negative' this prediction is actually correct in 90% of the cases (*negative predictive value*). Overall, the classifier is correct in 80% of the cases (*accuracy*) and the class 'positive' occurs in 50% of the observations (*prevalence*).

Some of these rates – like the false positive rate – are desired to be low, while others – like accuracy – are desired to be high. Generally, one can opt to maximise or minimise any of the above metrics and compare different performances of algorithms considering this metric. A frequently used performance indicator is *accuracy*, which shows how often the algorithm is correctly predicting a class. To evaluate if the accuracy of an algorithm is high or low, the R package 'caret' uses the so-called *No Information Rate* (NIR). The NIR can be considered as a naive classifier that just predicts the majority class for each data point. In the example above, the NIR would classify all observations as 'yes' since there are more 'yes' (275) than 'no' (207) observations in the data. Consequently, the NIR would correctly predict 275 out of all 482 cases or 57%. The accuracy of an algorithm can now be compared against the NIR and it can be statistically tested whether the accuracy of a classifier is significantly better than the NIR, which is obviously a desired outcome (Kuhn 2019, Chapter 17.2).

Evaluating an algorithm's performance using its accuracy is not unproblematic, especially when considering *unbalanced* datasets. In the machine learning terminology, *unbalanced* means that the proportions in the levels of the outcome are not equally distributed. If the outcome of interest, for example 'positive', is a rather rare phenomenon and is not observed in the dataset very frequently, while the outcome 'negative' is dominant, the dataset can be considered to be unbalanced. Unbalanced datasets are very common: illustrative examples here include detecting rare phenomena like fraud or diseases. There is no specific threshold for when a dataset is considered 'too unbalanced' to be valuable. Yet, it is important to check if a dataset is unbalanced or not to make informed decisions about which evaluation metrics to use.

The following example helps us to understand why an unbalanced dataset can be a problem for the accuracy as an evaluation metric: if a dataset with 100 observations includes only five observations with the outcome of interest, this can be considered an unbalanced dataset. Consider a complex algorithm, which achieved an accuracy of 87%, which means that this algorithm predicts 87 cases correctly. A naive predictor like the NIR, however, would instead simply predict that the observation does not have the outcome of interest (majority class) and it would perform even better than the accuracy of the complex algorithm since it would achieve a value of 95%, but at the cost of being completely incorrect in 5% of the cases because it was actually not able to predict anything else than 'does not have the outcome of interest'. As a result, accuracy is not an adequate choice as an evaluation metric when it comes to unbalanced datasets.

Due to this and other downsides, there are other related and more frequently used performance indicators, which can be derived from the confusion matrix. One of these indicators is the so-called *Receiver-Operator-Characteristic* (ROC). The ROC curve is a graphical illustration to evaluate the performance of a classifier by plotting the sensitivity against the false positive rate (or $1 - sensitivity$) and shows the trade-off between these two. An example is given in Figure 3. Sensitivity (true positive rate) is plotted on the horizontal axis while the false positive rate is plotted on the vertical axis. The curve is the performance of a discrete classifier, which predicts exactly one pair of sensitivity and false positive rate based on randomly generated data. A dashed 45° line represents those points that lie on exactly 50% sensitivity and 50% false positive rate. That means that any classifier on this line would predict exactly 50% of the sample correctly and the other half incorrectly. In other words, the prediction would be completely at random, comparable to a fair coin toss. On the other hand, a predictor hugging the top-left corner would have 100% sensitivity and 0% false positive rate meaning that all predictions would be correct, while any classifier below the dashed 45° line approximating the bottom-right corner would perform even worse than a guess and would misclassify more often than correctly predicting them.

It is desirable to find a classifier as close to the top-left corner as possible, undesirable to have an algorithm that follows or even undercuts the 45° line. Typically, the closer a classifier lies to the top-left corner, the better it gets. Thus, if two classifiers are compared, the one closer to the point $(1,1)$ is considered to perform better. One value that makes use of this characteristic is the Area Under the Curve or AUC. The AUC is simply the integral below a given classifier's curve. The more area a classifier can claim in the two-dimensional space – i.e., the closer it gets towards the top-left corner – the better the performance. Since the dashed 45° line cuts the space in half, its AUC value is 0.5 or 50%. The AUC of a classifier that hugs the top-left corner on the other hand would have a value of 1 or 100%. Thus, the desired value range of the AUC values for a well-performing classifier is greater than 0.5 and as close as possible or equal to 1. By comparing the AUC values of different curves, one can simply choose the curve with the highest AUC value as the one with the best performance (Fawcett 2006).

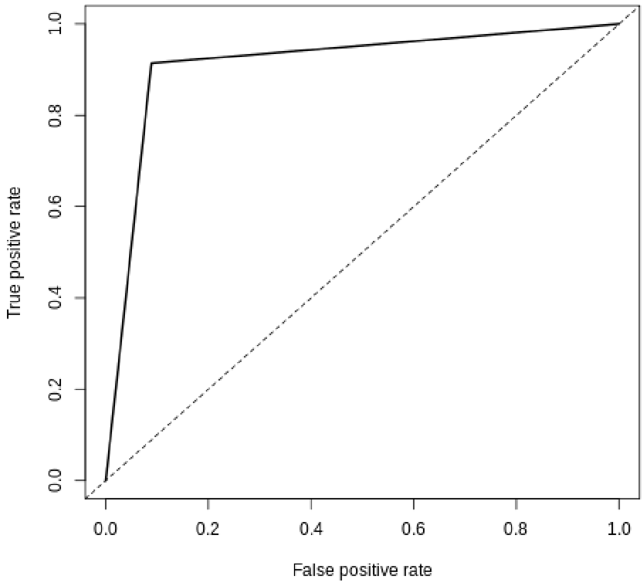

Figure 3: ROC Curve Analysis for Hypothetical Data

3.9.6. Algorithms Used in Empirical Chapters

The previous sections introduced the reader to most necessary and useful concepts and approaches to understand how machine learning works. Equipped with this background knowledge in data science, the reader will now be familiarised more with the algorithms that will be used in the later analyses of this thesis. There are a vast number of different (machine learning) algorithms that have been developed and enhanced during the last 80 years. Even though some of these techniques were already presented decades ago, only the computational power of today's machines enables researchers to apply them to the required large number of observations. In June 2020, the caret package for R was able to deploy 238 different machine learning algorithms (Kuhn 2019). From these, five are chosen to be compared in this thesis.

As a first algorithm, random forests are chosen. Random forests (RF or Forest) are based on decision trees, a technique widely known even in non-data related fields. They are commonly used and known to produce viable results in various contexts. Extreme Gradient Boosting (XGB) is chosen as the second alternative, often considered as one of the top performing algorithms over a wide range of applications. XGB is a method based on decision trees that improves the speed and predictive performance of random forests by optimising a technique called boosting. Next, Support Vector Machines (SVMs) are another very prominent algorithm widely applied and known to produce valuable predictions even (or especially) in highly complex data. In contrast to the before-mentioned methods, SVMs are not tree-based and can, thus, be seen as the direct rival of the XGB in this comparison. Since one research question of the later analyses is to shed light on whether applying complex machine learning algorithms on social science data is rewarding in a survey methodology context compared to using traditional approaches, these algorithms are compared to predictions from a binary logistic regression model (LOGIT). Lastly, to make further comparisons the RF, XGB and SVM are not only compared to predictions from LOGIT but also the LOGIT predictions are compared to its machine learning relative ‘Lasso and Elastic-Net Regularised Generalised Linear Models’ (GLMNET).

The following sections introduce the five algorithms and their applications for classification problems, to give the reader sufficient understanding of the algorithms and how they work on a practical basis, following up on the knowledge of the previous sections. The mathematical and technical details of the algorithms are explained as a part of this thesis here but are described extensively in the data science literature (Chen and Guestrin 2016; Kubat 2017; Müller and Guido 2016; Suthaharan 2016).

3.9.6.1. Random Forests

The underlying concept of random forests are the decision trees, which is why it is useful to explain the concept of decision trees first. A decision tree can predict both quantitative as well as qualitative responses and is referred to as a ‘regression tree’ in the former and as a ‘classification tree’ in the latter case, which will be the focus here. A classification tree tries to split observations of a dataset in such a way that the most similar cases are all assigned into the same group, called leaves, leaf-nodes or terminal nodes. If observations in a leaf are as similar as possible – meaning that their variance is low – the leaf is considered to be *pure*. This purity can be measured using different purity metrics, for example, the Gini coefficient, which measures the inequality in a frequency distribution. A low Gini coefficient indicates that the node predominantly contains observations from one single class and that the predictions were pure (James et al. 2013, p. 212).

When a tree is grown, the classification algorithm chooses the first single variable that produces the purest split into two subsets of observations as a starting point. If the chosen variable is categorical, it divides the data into two groups. If the chosen variable is continuous, the distribution is split in half at the best cut-off point to produce two sets of data (two starting leaves) that are as homogeneous within the leaves as possible. From there on, the algorithm looks for the next possible split in variables that further subdivides groups into their purest subgroups. If this process is done continuously, the tree would grow until it ends up with one terminal node for each observation and consequently a node size of one. Such a tree would yield perfect predictions for each training data point, but it is obviously of no use for test data because it would hopelessly over-fit the data. Therefore, the minimum node size is one of multiple hyperparameters that can be set before growing a tree. Once this the node size is set, a large tree will be grown, whose size is limited by the minimum node size.

Afterwards, this complex tree is *pruned* (for example, using the so-called *cost complexity pruning*). In a cross-validation process, the algorithm prunes the large tree and builds a subset of smaller trees that are as small as possible while still having the highest possible fit on the training data, and gives out the single best subtree for classification (James et al. 2013, p. 308). While single decision trees (both for regression and classification) have major advantages such as an easy and even visual interpretability, they tend not to have a high predictive power, which is why they are often enhanced using methods like Bootstrap Aggregation (or simply called *bagging),* boosting or through random forests, which are of interest in this thesis. In bagging approaches, instead of relying on a single decision tree, multiple samples are drawn with replacement (thus, the term ‘bootstrapping’) and new decision trees are trained for each sample. This process can be repeated dozens of times and at the end, the results will be aggregated to cancel out the downsides of single trees.

Moving from a single decision tree to random forests, on the other hand, it has to be noted that random forests combine multiple decision trees that are completely different in the way they are grown. Instead of training multiple identical decision trees on a large number of differently drawn samples with replacement (as in bootstrapping), random forests start by choosing a random subset of variables from the full variable set and choose only one of these variables to base the split decision on. From there on, it will continue to grow the tree based on an outcome criterion like purity. As the term ‘forest’ hints, many trees are grown using this approach – all starting at a random cut-off point of a variable. The ingenuity lies in the fact that a regularly grown decision tree would always choose the variable with the best possible split as the starting point. If multiple decision trees were grown using the full predictor set, the decision trees would all start to split the data using the same most influential predictor. By only allowing the algorithm to base the decision on variables from a randomly chosen subset of predictors, chances are that the influential predictor is not even part of this subset and, thus, the algorithm is forced to grow a completely different tree. This technique leads to less correlation between trees compared to trees that are all grown based on bootstrapped training data. Eventually, the outcomes of the different trees (or the forest) are averaged to create one predictive algorithm that is less variable and shows higher predictive power than one single, high variance decision tree. The advantage of random forests over decision trees lies in the highly improved accuracy of its predictions. However, random forests cannot be illustrated as easily as a single tree and the interpretability is, thus, less straightforward (James et al. 2013, p. 320).

3.9.6.2. eXtreme Gradient Boosting

Besides bagging and random forests, boosting techniques are another approach to get more information from simple decision trees and are consequently most commonly, but not exclusively, seen in the context of regression or classification trees. In contrast to bagging, boosting does not rely on bootstrapping and does not grow multiple trees from bootstrapped samples. Instead, trees in boosting approaches are grown step by step using information from previously grown trees (James et al. 2013, p. 223). Rather than immediately trying to predict the target variable, boosting approaches try to fit smaller models on the residuals of a given (baseline) model. In the next iteration, the residual predictions are fed back into the model and the model is re-trained using these residuals as weights to re-estimate the residuals. Using this iterative approach, the model is slowly learning how to improve the prediction in areas where it was not performing well before. Extreme Gradient Boosting (XGB or XGBoost) is meant to ‘push the extreme of the computation limits of machines to provide a scalable, portable and accurate library’ (XGBoost Developers 2020) to speed up machine learning processes. XGB was particularly designed to control over-fitting better by leveraging the bias-variance trade-off and achieving a better performance by making use of parallel computing. In the caret package, XGB is integrated with the ‘xgboost’ package (Chen et al. 2020).

3.9.6.3. Support Vector Machines

Support Vector Machines (SVM) are well established and were developed many years ago. However, due to their complexity, they tended not to be feasibly deployable due to a lack of computational power. This has recently changed with technological advances and nowadays SVMs are widely applied in a range of predictive analytics contexts. Social scientists and other researchers trained in traditional statistics might be reminded of Ordinary Least Squares (OLS) regression when looking at the basic principles of SVMs in a two-dimensional space or the maximal margin classifier, in particular. However, despite some conceptual parallels, SVMs and OLS regression serve very different purposes.

In a two-dimensional cloud of data-points, a maximal margin classifier tries to find the one separating line that classifies the cloud as best as possible into two distinct groups, while also maximising the perpendicular distance between the nearest point and the separating line, to prevent any misclassification. If lines are drawn through those points that have the maximal distance to the separating line and are running parallel to this separator, these new lines are referred to as ‘support vectors’. Support vectors determine the position of the maximal margin classifier in a sense that if they (or more precisely the data point they are based on) were moved, the line defining the maximal margin classifier would also move in the two-dimensional space (James et al. 2013, p. 338).

These same concepts also translate to higher dimensions (more than two variables): in a three-dimensional space the maximal margin classifier and support vectors describe hyperplanes, while in even higher (yet more realistic) dimensions the exact shape of the separator become unimaginable. Dealing with high dimensional data is not an obstacle for SVMs. However, with rising data complexity it becomes more difficult to find a linear separator. In fact, SVMs are particularly well equipped to deal with data that is non-linearly separable. While the mathematical and technical details of the so-called 'kernel-trick' are beyond the scope of this thesis, they are also clearly explained elsewhere (James et al. 2013, p. 350). The basic idea to cope with non-linearity is – again, like in the OLS regression – to enlarge the feature space using quadratic, cubic or higher-order polynomial functions or interaction terms of the predictors. The solution to a classification problem in an artificially enlarged feature space then leads to a linear decision boundary that makes classification of data points possible (James et al. 2013, p. 344-355).

The mathematical and technical complexity of SVMs is harder to understand than that of random forests and XGB, which makes it challenging for non-mathematicians to grasp and explain what exactly these algorithms do in detail. However, to distinguish the methods, it is important to realise that while the before-mentioned approaches are following the logic of decision trees, SVMs introduce a different logic of finding a decision boundary in the form of vectors or hyperplanes.

3.9.6.4. Lasso and Elastic-Net Regularised Generalised Linear Models

The ‘Lasso and Elastic-Net Regularised Generalised Linear Model’ or ‘glmnet’ package for R is a powerful collection of different algorithms, which can be used to fit linear, binary logistic, multinomial logistic, grouped multinomial, multiple-response Gaussian or Poisson regression as well as Cox models. The package glmnet is compatible with caret and can, thus, be trained using caret’s ‘train.control’ argument. The power of glmnet stems from its ability to efficiently determine combinations of the ‘Least Absolute Shrinkage and Selector Operator’ (LASSO) and the so-called ridge regression, as well as penalty parameters through cross-validation for generalised linear models. The idea behind both LASSO and ridge regression is to tune a model as best as possible. In the case of a ridge regression, this means reducing the influence of less influential variables as much as possible but never setting the influence exactly to zero. The LASSO extends this idea by allowing the influence of variables to equal zero. The process of reducing these influences is called ‘shrinking’ or ‘regularisation’ and for this reason, the two methods are summarised, alongside others, under the broader terms of ‘shrinkage’ or ‘regularisation’ methods. Furthermore, setting the coefficients equal to zero in the LASSO regression instead of only allowing them to approximate zero, like in the ridge regression, literally nullifies the impact of these variables on the outcome and basically excludes them. Thus, LASSO regression can be used to select a predictor subset of influential variables indicated by strictly non-zero coefficients. This makes LASSO regression models more parsimonious, reduces the likelihood of overfitting and increases interpretability, since it reduces model complexity alongside multicollinearity. In this light, LASSO regression can be considered an alternative to more familiar approaches in the social sciences like forward or backward selection (James et al. 2013, p. 2019ff).

While the technical and mathematical details can be found in Friedman, Hastie, and Tibshirani (2010), it is important to emphasise that glmnet balances a hyperparameter *alpha* to lie between 0 and 1. If *alpha* is equal to 0, a pure LASSO regression is performed. By contrast, if *alpha* is equal to 1 a pure ridge regression is carried out. Values between the endpoints indicate a combination of both approaches. Thus, glmnet tries to find the best trade-off between a smaller model while keeping as many variables as needed by only reducing their coefficients (Friedman et al. 2020).

For the analyses in this dissertation, the glmnet package is used to solve the classification problem of first contact attempt by using an efficiency-tuned logistic regression, which leverages the advantages of the ridge and lasso regressions. The objective of using this technique, which can be seen as a machine learning variant of a logistic regression, is to find out whether the algorithm can outperform the traditional logistic regression analysis when applied to the same data.

3.9.6.5. Logistic Regression

Besides the different machine learning algorithms discussed so far, a binary logistic regression without any machine learning features will be used to predict the success of the first contact attempt. Logistic regression is part of the quantitative toolbox that social scientists use and a typical model of choice if the outcome of interest is categorical, whether of a binary, ordinal or nominal nature. Because OLS regression or linear probability models have difficulties in estimating coefficients for categorical dependent variables e.g. running estimations outside the target variable's logical boundaries and heteroscedasticity (Kopp and Lois 2014, p. 162f; Menard 2002, p. 4f), logistic regression, which basically estimates the logarithm of the odds, becomes the go-to technique for categorical outcomes.

In the analyses of this thesis, a binomial logistic regression for a given variable set is estimated using the train data.

It is important to note that the regression model only includes main effects. Although a model which applies a more complex specification could prove useful, it was purposely not used because specifying a logistic regression model as best as possible is not the aim of this thesis. This implies a decision about when to stop specifying and how elaborated the specifications should be. After including interaction terms, a logical next step would be to model a multi-level structure and would inevitably involve a good deal of 'intervention' on the part of the researcher. Relying on an empirically driven specification search (without the division between train and test data sets) runs the risk of over-fitting a model and over-using chance patterns in the data; on the other hand, the literature review chapter demonstrates that research to date does not converge on a single theoretically derived model. Manually specifying only the logistic regression while leaving the specification of the machine learning approaches untouched could be judged to negatively affect and bias the comparison. Because of this it was decided to stick to a main effects model.

While this approach might be arguable it is still of scientific interest. In a sense it can be compared to David versus Goliath, with David being the logistic regression. The expected result is that machine learning algorithms outperform the ‘under-specified’ logistic regression model. In such a case it would be interesting to find out how much specification is needed to keep up with the machine learning algorithms. If, however, already the main-effects logistic regression was able to keep up or even outperform the machine learning algorithms, the feasibility of such complex algorithms could be deemed questionable.

The coefficients derived from the trainset model are used to predict the class of those observations assigned to the testset. An observation is classified as ‘contact’ if the estimated probability is equal to or greater than 0.5 and as ‘non-contact’ if it is below this threshold.

Investigating whether the predictive performance of a logistic regression can outperform predictions of more elaborated machine learning techniques will be one focus in the upcoming analyses.

In Chapter 3 the data and methods for the analyses of the later chapters have been introduced. It was described how the data is pre-processed, what the exclusion criteria were and how the correlates, which were derived from the literature, were operationalised and with which statistical methods they will be analysed. Moreover, the second objective of this chapter was to familiarise the reader with the fundamental concepts of data science and machine learning in Sections 3.8 and 3.9. This background information of the data and methods is important to be able to follow along and contextualise the analyses in the second part of this thesis. Chapter 3 can therefore be considered as a compendium to guide the reader through the empirical chapters of this thesis, which start in Chapter 4 by presenting the results from an analysis of the correlates of contact in the ESS Round 9.

PART II

4. Is Anybody There?

An Analysis of Contact Success in the ESS 9 in the United Kingdom, Germany, and France

The first part of the thesis provided the relevant background information and laid the foundation to understand the relevance, context, and state of research on contactability. In the second part the first contact attempt success will be analysed using different methods and datasets. While Chapters 4 and 5 will utilise ESS Round 9 data, Chapters 6 and 7 will draw on a pooled ESS dataset. While Chapters 5, 6 and 7 strongly rely on the machine learning methods introduced earlier in Section 3.9, a solid foundation for these advanced methods is first built in this chapter, which examines the correlates of contact drawing on the findings from the literature review. The chapter first emphasises the ESS Round 9 response rates in the UK, Germany and France since response rates are dependent on successful contact in the first place. Then it reports an analysis of the correlates of the first contact attempt, investigating a total of four high-level research questions. The overview of the operationalisation of variables (Table 2 in Section 3.3) which are investigated in this chapter might be a useful companion throughout the analyses.

Response rates in the United Kingdom, Germany and France are far off from the ESS' quality requirement of a 70% response rate. Figure 4 shows the overall survey response rates for each of the three countries for all nine ESS rounds[11]. The three countries have not met the target response rate since the start of the survey but while response rates for France seem to have increased in the recent years, with a slight positive trend and the highest in the ESS 9 compared to the other countries, response rates in the UK and Germany have declined rapidly. While the UK had a response rate of 55% in ESS 1, in ESS 9 this dropped down to 41%. More notably, response rates in Germany experienced a marked decline from 55% in ESS 1 to only 27% in ESS Round 9.

[11] Response rates definitions have changed over time and were slightly different in ESS 1 to 2 compared to ESS 3 to 8, whereas the response rate definition was not available for ESS 9. Details can be found in the survey documentation reports for each wave (see, for example, European Social Survey 2018c).

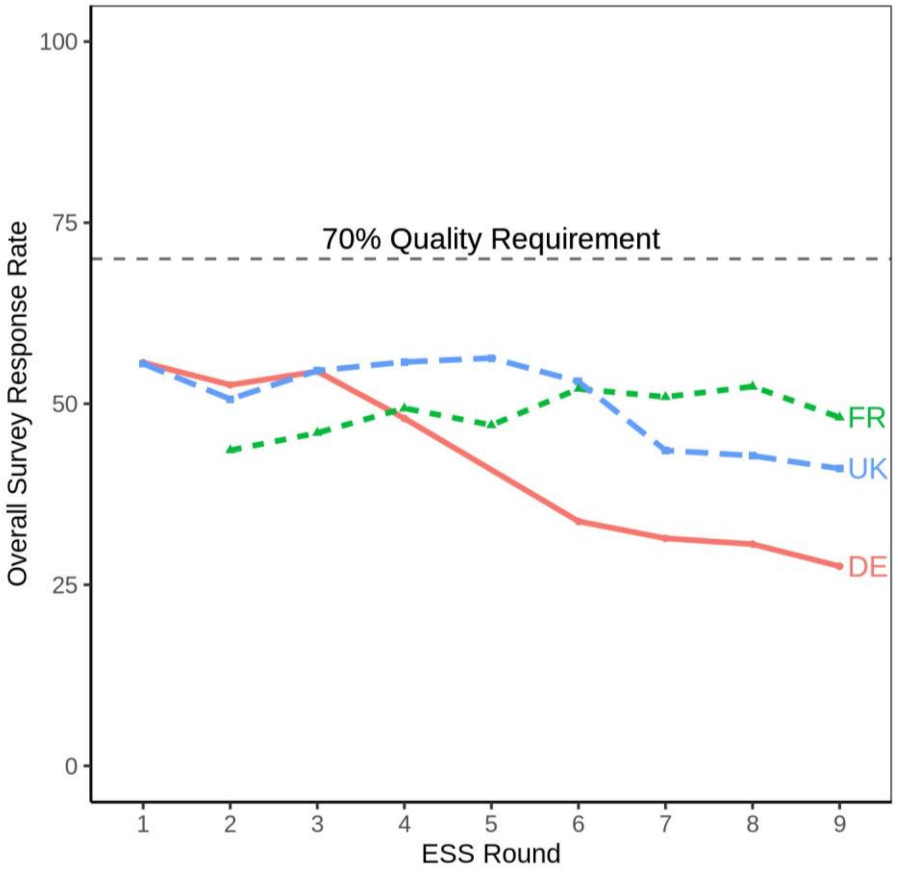

Figure 4: Response Rates of the Three Example Countries from ESS 1 to ESS 9.

Investigations into any form of unit nonresponse is of crucial interest, especially to better understand its mechanisms and characteristics to eventually improve and increase response rates. As noted in Section 2.1, nonresponse and contactability are discussed widely in the literature together, as contact success is a prerequisite for unit response or nonresponse. This chapter focuses on investigating the correlates of a successful contact for the first contact attempt in the ESS Round 9. This not only helps understand contact success in the ESS but also supports efforts in understanding nonresponse.

4.1. Research Questions

The following section is organised country by country, and in each case proceeds by examining four research questions. These relate to characteristics of the fieldwork, area and household, interviewers, and potential respondents. In detail, the following research questions will be answered:

Research Question 1: Is a successful first contact attempt associated with the day of the week and time of the day?

> To investigate the characteristics that are related to the fieldwork procedures and the contact attempt characteristics it will be examined whether or not there is an association between a first contact attempt success and the chosen day of the week or time of day.

Research Question 2: With which area or household characteristics is first contact attempt success associated?

> To investigate area and household level attributes it, will be explored whether the degree of urbanicity, the type of house a unit lives in, the building's physical condition as well as the presence/absence of access impediments, the degree of decay indicated by rubbish/litter or vandalism/graffiti in the area, or whether or not a telephone number is available for the target unit is associated with first contact attempt success.

Research Question 3: With which interviewer characteristics is first contact attempt success associated?

With respect to the interviewer level, it will be explored whether an interviewer's workload, completion rate, sex or age is associated with first contact attempt success.

Research Question 4: With which potential respondent characteristics is first contact attempt success associated?

Lastly, considering the potential respondents characteristics it will be investigated whether the potential respondent's sex[12], the presence/absence of own children at home, marital status, household size, employment status, education, household income or age is associated with a first contact success.

4.2. Results

After downloading and merging, the UK dataset for the European Social Survey Round 9 included 5,850 observations, the German dataset 8,695 and the French dataset 4,400 observations on 1,502 variables. Of all 5,850 observations in the UK, 2,204 of those units participated in the survey. For the remaining 3,646 UK non-participating observations, only data from the contact information dataset is available. In the German dataset, 2,358 of 8,695

[12] The survey literature often uses interviewer/respondent sex or gender as interchangeable terms even though they are distinct socio-demographic and identity concepts. This dissertation only examines 'sex' as a socio-demographic characteristic of the interviewer or the respondent.

observations participated in the survey, and in the French sample, of all 4,400 observations there were 2,010 participants. Following the exclusion criteria (see Section 3.2) for this study, 360 from 5,850 UK units were excluded, 224 units were excluded in the German and 139 excluded in the French sample. This leads to a net-sample size of 5,490 observations in the UK sample (2,165 respondents, 3,320 nonrespondents), 8,471 observations in the German sample (2,353 respondents, 6,118 nonrespondents) and 4,261 observations for the French sample (2,010 respondents, 2,251 nonrespondents). The following flow-chart illustrates the data selection beginning with the entire dataset of 30 countries that are available in ESS Round 9.

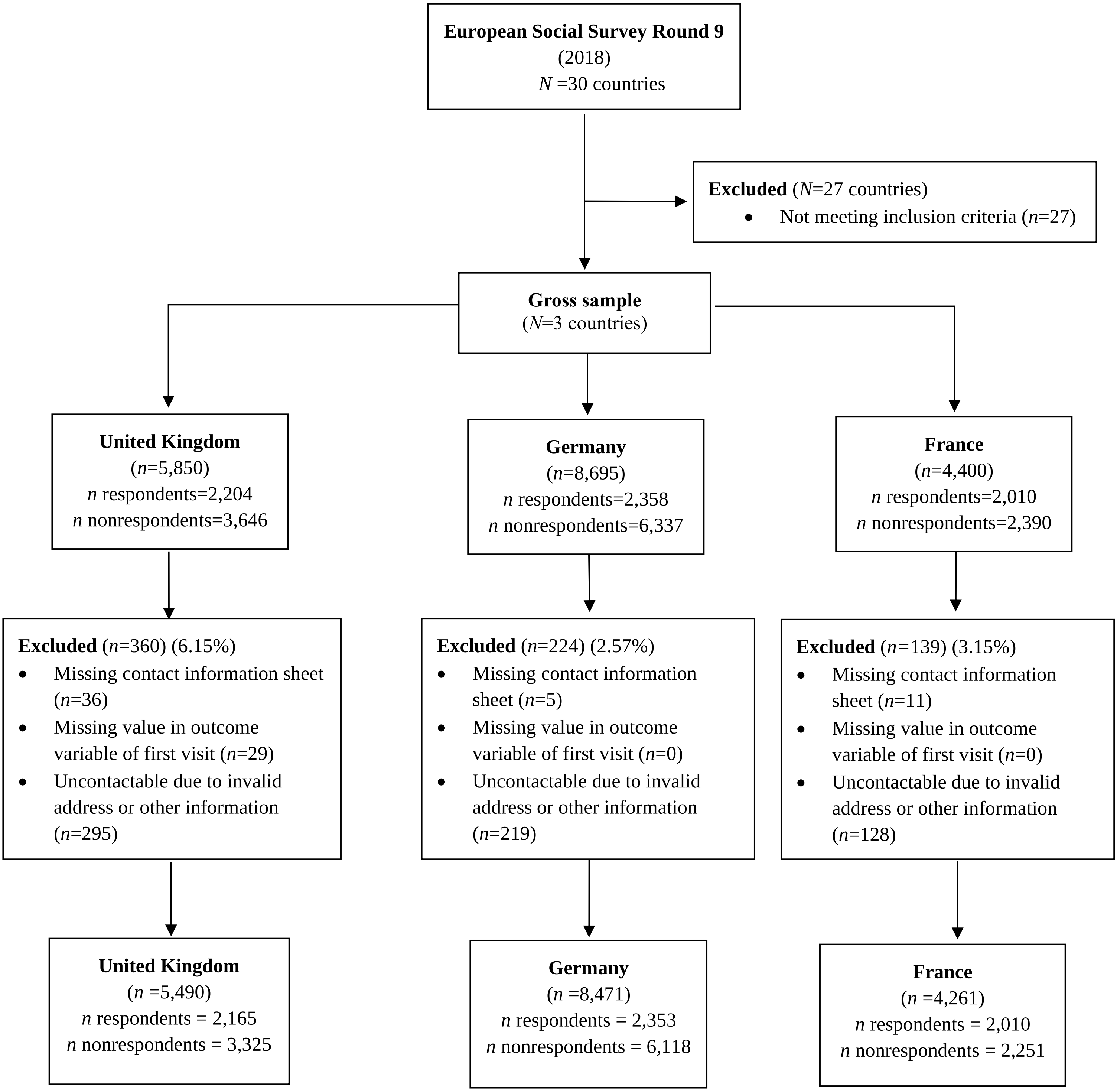
European Social Survey Round 9
(2018)
N =30 countries
Excluded (*N*=27 countries)
• Not meeting inclusion criteria (*n*=27)
Gross sample
(*N*=3 countries)
United Kingdom
(*n*=5,850)
n respondents=2,204
n nonrespondents=3,646
Germany
(*n*=8,695)
n respondents=2,358
n nonrespondents=6,337
France
(*n*=4,400)
n respondents=2,010
n nonrespondents=2,390
Excluded (*n*=360) (6.15%)
• Missing contact information sheet (*n*=36)
• Missing value in outcome variable of first visit (*n*=29)
• Uncontactable due to invalid address or other information (*n*=295)
Excluded (*n*=224) (2.57%)
• Missing contact information sheet (*n*=5)
• Missing value in outcome variable of first visit (*n*=0)
• Uncontactable due to invalid address or other information (*n*=219)
Excluded (*n*=139) (3.15%)
• Missing contact information sheet (*n*=11)
• Missing value in outcome variable of first visit (*n*=0)
• Uncontactable due to invalid address or other information (*n*=128)
United Kingdom
(*n* =5,490)
n respondents = 2,165
n nonrespondents = 3,325
Germany
(*n* =8,471)
n respondents = 2,353
n nonrespondents = 6,118
France
(*n* =4,261)
n respondents = 2,010
n nonrespondents = 2,251

Figure 5: Flow Chart of Data Selection Process.

Table 5 shows the sample composition of the fieldwork-related variables compared across the three countries under investigation. France achieved more than 8.5% of completed or at least partially completed interviews at the first contact attempt. In Germany, on the other hand, completed or partial interviews were only possible for 1.2% of first contact attempts. Finally, the UK falls somewhere in between (4.8%). In the UK 22.4%, in Germany 30.7% and in France 26.1% of contacts were made with persons other than the respondent or with unidentified persons. Even if contact with the potential respondent was established, it did not lead to an interview in the German sample, more often so than in the samples in the UK or France (UK: 16.5%, Germany: 35.1%, France: 7.2%). The regrouping of these outcomes shows that a successful first contact was established in the UK in 43.8% of the cases and in France in 41.7% of the cases, whereas Germany showed a higher successful first contact attempt rate of 67.0%.

In the UK, first contact attempts were fairly evenly distributed over the weekdays and Saturdays with first contact attempts most often made on Tuesdays (21.2%) and least often on Sundays (5.7%). A similar but even more uniform distribution of contact attempts from Mondays to Saturdays can be seen for Germany, where first contact attempts most often took place on Saturdays (18.3%). Similarly, to the UK, Sundays were also rarely chosen for first contact attempts by the interviewers (2.9%). The French data shows a different pattern. While there were few first contact attempts during the week from Monday to Friday, almost half of all first contact attempts were made on Saturdays (48.8%) and none on Sundays. In terms of contact hour, on average, first contact attempts were made earliest in the UK, followed by Germany and France with the latest average first contact hour.

	United Kingdom		**Germany**		**France**	
	n	**%**	**n**	**%**	**n**	**%**
Result of first visit *(n)*	5,490	100.00	8,471	100.00	4,261	100.00
Completed or partial interview	265	4.83	104	1.23	360	8.45
Contact with respondent but no interview	905	16.48	2,975	35.12	305	7.16
Contact with other than respondent	1,232	22.44	2,596	30.65	1,113	26.12
No contact at all	3,088	56.25	2,796	33.01	2,483	58.27
missing	0	-	0	-	0	-
First contact success *(n)*	5,490	100.00	8,471	100.00	4,261	100.00
No	3,088	56.25	2,796	33.01	2,483	58.27
Yes	2,402	43.75	5,675	66.99	1,778	41.73
missing	0	-	0	-	0	-
Day of week for first visit *(n)*	5,489	100.00	8,471	100.00	4,261	100.00
Monday	1,052	19.17	1,466	17.31	465	10.91
Tuesday	1,162	21.17	1,449	17.11	454	10.65
Wednesday	937	17.07	1,389	16.40	474	11.12
Thursday	731	13.32	1,349	15.92	366	8.59
Friday	714	13.01	1,020	12.04	422	9.90
Saturday	579	10.55	1,551	18.31	2,080	48.81
Sunday	314	5.72	247	2.92	0	0.00
missing	1	-	0	-	0	-
Hour of day for first visit *(n)*	5,488	100.00	8,471	100.00	4,261	100.00
mean (sd)	14.42	(2.27)	15.26	(2.69)	15.73	(3.05)
missing	2	-	0	-	0	-

Table 5: Descriptive Fieldwork Data by Country. (n) Represents Valid Cases.

Figure 6 shows density plots for the first contact hour in all three countries, with their means, all around 3pm. Looking at the country differences, the graph shows that the plots for Germany and France are both skewed to the left indicating the relatively more frequent practice of later contact hours while the UK sample is almost symmetrical with a slight positive skew, indicating that earlier first contact hours were relatively a little more common.

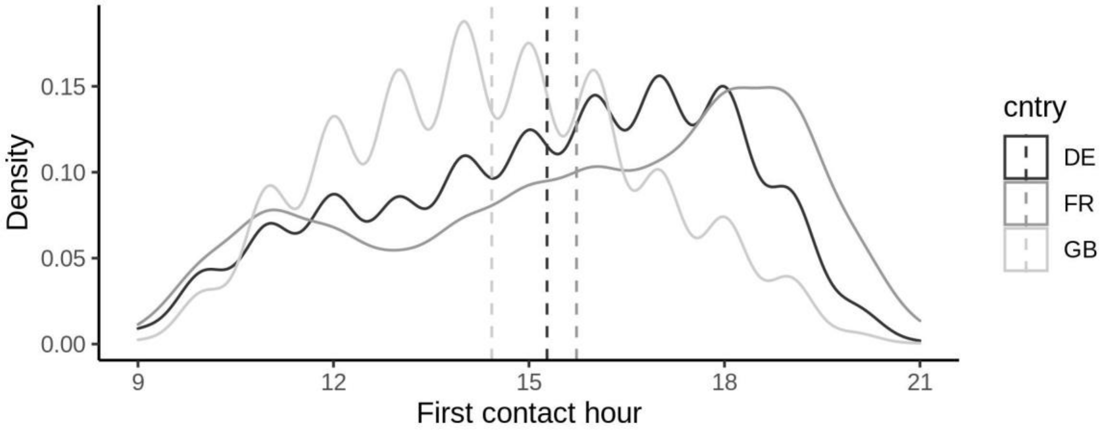

Figure 6: Density Plots of First Contact Hour by Country

Figure 7 shows heatmaps for the three countries combining the day of the first contact (*y*-axis) with the first contact hour (*x*-axis). Each cell, thus, represents a specific day-time combination throughout the week. The frequency of how many first contact attempts were made at every specific day-time combination is indicated by the colour code, where darker values represent a higher frequency of attempts. In the UK, first contact attempts happened predominantly in the early week between 1pm and 4pm with contacts most likely to be made on Mondays at around 2pm. In Germany first contacts were mostly made from Mondays to Thursdays from around 4pm to around 7pm and Saturdays from around 11am to around 4pm with contact attempts notably concentrated on Mondays, Tuesdays and Thursdays around 6pm as well as Wednesdays around 5pm. The heatmap for the French sample shows that there were few weekday contacts happening, mostly between around 6pm and around 7pm where weekdays were used. Most of the first contact attempts in France were made on Saturdays between 10am and 5pm, peaking at around 11am and 3pm.

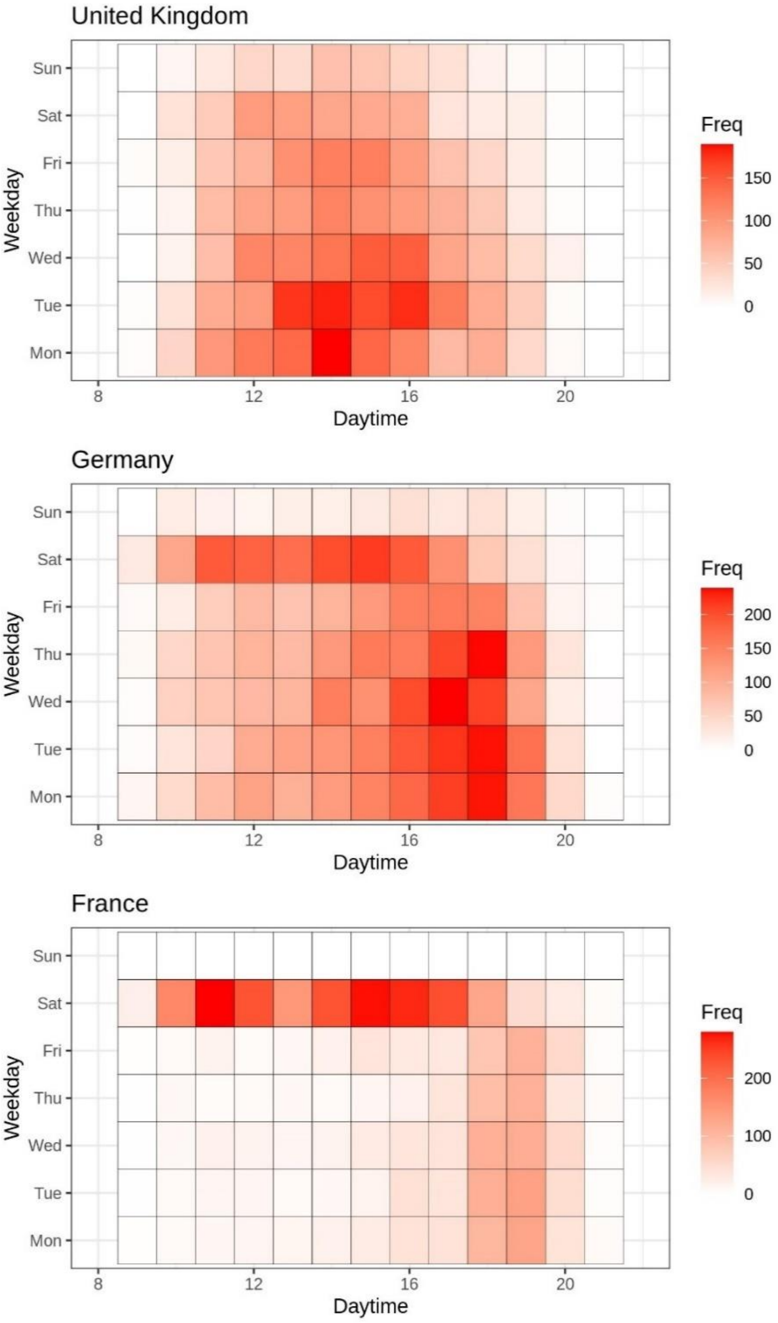

Figure 7: Heatmaps for First Contact Hour by Day of the Week

Table 6 shows the sample composition for the housing and area related variables for the three countries under examination. While 3.6% of the sample in the UK and 2.5% of the sample in Germany describe their domicile as being a farm or a home in the countryside, 6.5% of the sample in France consider themselves as living in the same settings. By contrast, while roughly a third of German (35.8%) and French (34.7%) respondents categorise their home as being in a town or small city, 45.7% of UK respondents report themselves belonging in this category. Considering the type of house the respondent lives in, results show that there was a larger proportion of single-unit houses in the UK sample (79.5%) than in the German (51.8%) or French sample (58.6%). While 'other' types of houses are rare in all country samples (UK: 1.9%, Germany: 0.4%, France: 0.6%), multi-unit housing is much less prominent in the UK sample (18.6%) compared to the German (47.8%) and French (40.8%) samples.

Most housing conditions in the French sample were described as 'good or very good' (82.8%), while this compares to 71.0% in Germany and 68.1% in the UK. Both the amount of rubbish and litter or graffiti and vandalism in the immediate vicinity were similar across all countries. Germany showed the highest proportion of rubbish in the immediate vicinity with 4.6% of the cases, while the French sample featured the lowest proportion (2.0%). The results for access impediments are mixed: while only 14.8% of the buildings of the UK sample units were equipped with access impediments, 49.5% of French addresses and 74.2% of German addresses had such features. Further differences can be found when investigating whether telephone numbers were present for the sample members or not. While for France no data on telephone numbers is available at all, data on telephone numbers in the UK was only collected from a small proportion of respondents. In Germany, telephone numbers were available even if the unit did not respond to the survey and were available for 23.9% of units.

	United Kingdom		**Germany**		**France**	
	n	**%**	**n**	**%**	**n**	**%**
Domicile, respondent's description *(n)*	2,165	100.00	2,353	100.00	2,009	100.00
Farm or home in countryside	77	3.56	59	2.51	131	6.52
Country village	464	21.43	713	30.30	557	27.73
Town or small city	990	45.73	843	35.83	697	34.69
Suburbs or outskirts of big city	441	20.37	376	15.98	277	13.79
A big city	193	8.91	362	15.38	347	17.27
missing	3,325	-	6,118	-	2,252	-
Type of house respondent lives in *(n)*	4,872	100.00	7,909	100.00	4,207	100.00
Single unit	3,873	79.50	4,100	51.84	2,465	58.59
Multi-unit	905	18.58	3,777	47.76	1,716	40.79
Other	94	1.93	32	0.40	26	0.62
missing	618	-	562	-	54	-
Physical condition of building/house *(n)*	4,822	100.00	7,906	100.00	4,261	100.00
Bad or very bad	169	3.50	349	4.41	122	2.86
Satisfactory	1,368	28.37	1,944	24.59	611	14.34
Good or very good	3,285	68.13	5,613	71.00	3,528	82.80
missing	668	-	565	-	0	-
Access impediments *(n)*	4,804	100.00	7,868	100.00	4,261	100.00
No access impediments	4,093	85.20	2,027	25.76	2,153	50.53
Access impediments	711	14.80	5,841	74.24	2,108	49.47
missing	686	-	603	-	0	-
Litter in immediate vicinity *(n)*	4,829	100.00	7,911	100.00	4,261	100.00
None, almost none or small amount	4,727	97.89	7,548	95.41	4,175	97.98
Large or very large amount	102	2.11	363	4.59	86	2.02
missing	661	-	560	-	0	-
Vandalism in immediate vicinity *(n)*	4,829	100.00	7,911	100.00	4,261	100.00
None, almost none or small amount	4,808	99.57	7,743	97.88	4,201	98.59
Large or very large amount	21	0.43	168	2.12	60	1.41
missing	661	-	560	-	0	-
Telephone number available *(n)*	1,890	100.00	8,471	100.00	0	0.00
No	1,851	97.94	6,444	76.07	0	0.00
Yes	39	2.06	2,027	23.93	0	0.00
missing	3,600	-	0	-	4,261	-

Table 6: Sample Composition for Housing and Area Variables. (n) Represents Valid Cases.

Table 7 shows the sample composition for the included interviewer-related variables for the three countries under examination. Some 1,873 of the 2,165 UK respondents (86.5%) were interviewed by the same interviewer who reached out for the first contact attempt. In Germany, this proportion reaches 94.4% and in the French sample even 96.8% of the respondents were interviewed by the same interviewer who made the first contact attempt. Overall, 284 unique interviewer numbers were found in the UK sample, 211 in the German and 195 in the French one. While the sex distribution of the interviewers was fairly even in the UK sample (51.3% female interviewers), an imbalance in the German sample was found (41.3% female) and an even more distinctive disparity was observed in the French sample with 64.4% female interviewers. Interviewers tended to be older on average in the German sample (*mean*=61.5, *sd*=10.5), followed by UK interviewers (*mean*=58.0, *sd*=10.4) while French interviewers were the youngest on average (*mean*=53.3, *sd*=10.0). The average first contact workload was highest in Germany (*mean*=40.1, *sd*=20.0), whereas the average first contact attempt workload was similar in the UK (*mean*=19.3, *sd*=9.4) and France (*mean*=21.9, *sd*=10.8). The highest average interviewer completion rate, as operationalised in Section 3.3., can be found in the French sample (*mean*=48.4, *sd*=19.8), followed by the UK sample (*mean*=39.8, *sd*=19.9) and the German sample (*mean*=26.8, *sd*=12.4). However, when looking at the first contact success rate, the picture is different: the highest first contact success rate can be found in Germany (*mean*=64.4, *sd*=24.0), followed by the UK (*mean*=44.3, *sd*=18.2) and France (*mean*=43.2, *sd*=20.4).

	United Kingdom		**Germany**		**France**	
	n	**%**	**n**	**%**	**n**	**%**
Unique first contact interviewer IDs	284	100.00	211	100.00	195	100.00
First contact interviewer also final interviewer? *(n)*	2,165	100.00	2,353	100.00	2,010	100.00
Yes	1,873	86.51	2,221	94.39	1,945	96.77
No	292	13.49	132	5.61	65	3.23
missing	3,325	-	6,118	-	2,251	-
Workload of first contact interviewer *(n)*	284	100.00	211	100.00	195	100.00
mean (sd)	19.33	(9.40)	40.14	(19.97)	21.85	(10.78)
missing	0	-	0	-	0	-
Completion rate *(n)*	278	100.00	206	100.00	191	100.00
mean (sd)	39.80	(19.85)	26.83	(12.36)	48.36	(19.78)
missing	6	-	5	-	4	-
First contact success rate *(n)*	284	100.00	211	100.00	195	100.00
mean (sd)	44.26	(18.22)	64.43	(23.96)	43.19	(20.38)
missing	0	-	0	-	0	-
Sex of first interviewer *(n)*	277	97.53	206	97.63	191	97.94
Female	142	51.26	85	41.26	123	64.40
Male	135	48.74	121	58.74	68	35.60
missing	0	-	0	-	0	-
Age of the first interviewer *(n)*	277	97.53	206	97.63	191	97.94
mean (sd)	58.02	(10.36)	61.48	(10.50)	53.25	(10.04)
missing	0	-	0	-	0	-

Table 7: Sample Composition for Interviewer Data. (n) Represents Valid Cases.

Table 8 shows the sample composition of the respondent-related variables for the three countries under examination. While the sex of sampled units is likely closer to the population distribution in the German sample (48.7% female), the samples in the UK and France show a slight surplus of females with roughly 55% in both. The samples feature comparable proportions of units with children at home (UK: 46.1%, Germany: 46.0%, France: 44.9%). While the German and French samples contain similar proportions of married respondents (4.8% and 4.1%, respectively), the UK sample consists of 11.8% married units. The highest share of separated or divorced units can be found in the French sample (24.8%), while the highest proportion of people who do not fall in the married, separated/divorced or widowed categories can be found in Germany (64.3%). About half of the French sample (47.5%) were in paid work during the last seven days. This compares with 51.0% of the German and 52.4% of the UK sample who were in paid work. The highest share of unemployed sampled units can be found in the French sample (5.5%). The French sample also features a higher proportion of retired sample members (34.1%) compared to the UK (28.4%) and Germany (24.0%). The average household size of sampled units was roughly the same across all three countries (UK: *mean*=2.3, *sd*=1.3; Germany: *mean*=2.6, *sd*=1.3; France: *mean*=2.2, *sd*=1.3). On average, the UK sampled units had 14 years and 80 days of education, German units 14 years and 105 days of education and French units 13 years and 43 days of education. The highest average household income decile was found in the German sample, followed by the UK and French samples. While units in the UK and French samples are roughly of the same mean age (52), units from the German sample are about three years younger on average.

	United Kingdom		**Germany**		**France**	
	n	**%**	**n**	**%**	**n**	**%**
Respondent's sex *(n)*	2,165	100.00	2,353	100.00	2,010	100.00
Female	1,186	54.78	1,146	48.70	1,097	54.58
Male	979	45.22	1,207	51.30	913	45.42
missing	3,325	-	6,118	-	2,251	-
Children at home *(n)*	1,204	100.00	1,058	100.00	1,101	100.00
No	649	53.90	571	53.97	607	55.13
Yes	555	46.10	487	46.03	464	44.87
missing	4,581	-	7,632	-	3,288	-
Respondent's marital status *(n)*	1,225	100.00	1,107	100.00	1,156	100.00
Married	144	11.76	53	4.79	47	4.07
Separated/divorced	257	20.98	191	17.25	286	24.74
Widowed	209	17.06	151	13.64	213	18.43
None of these	615	50.20	712	64.32	610	52.77
missing	4,265	-	7,364	-	3,105	-
Main activity in the last 7 days *(n)*	2,163	100.00	2,347	100.00	2,008	100.00
Paid work	1,134	52.43	1,198	51.04	954	47.51
Education	75	3.47	226	9.63	121	6.03
Unemployment	61	2.82	60	2.56	110	5.48
Sick	94	4.35	53	2.26	71	3.54
Retired	614	28.39	562	23.95	685	34.11
Housework	172	7.95	203	8.65	57	2.84
Other	13	0.60	45	1.92	10	0.50
missing	3,327	-	6,124	-	2,253	-
Number of household members *(n)*	2,178	100.00	2,351	100.00	4,261	100.00
mean (sd)	2.32	(1.27)	2.58	(1.27)	2.23	(1.27)
missing	3,325	-	6,120	-	2,251	-
Years of education *(n)*	2,162	100.00	2,346	100.00	4,261	100.00
mean (sd)	14.22	(3.73)	14.29	(3.49)	13.12	(4.17)
missing	3,341	-	6,125	-	2,289	-
Household income decile *(n)*	1,829	100.00	2,346	100.00	1,791	100.00
mean (sd)	5.18	(2.97)	6.07	(3.49)	4.99	(3.04)
missing	3,671	-	6,125	-	2,470	-
Respondent's age *(n)*	2,150	100.00	2,349	100.00	2,010	100.00
mean (sd)	52.43	(18.43)	49.66	(19.06)	52.37	(18.97)
missing	3,340	-	6,122	-	2,251	-

Table 8: Sample Composition for Respondent Data. (n) Represents Valid Cases.

Table 9 shows cross-tabulations for the outcome variable of the share of successful first contacts and the chosen weekday for the first contact attempt. In the UK, the most successful first attempts were made on Sundays with 52.2% of attempts made on this day being successful. The least successful day for first contact attempts were Fridays with 40.9% successful contacts.

A χ^2-test of independence returned a statistically significant result, suggesting that these two variables are not independent of each other ($\chi^2 = 22.1$, $df = 6$, $p = 0.0012$). In Germany, Mondays proved to be the most successful days for establishing first contact, resulting in 70.9% of contacts being successful. The lowest rate of first contact success was obtained on Thursdays (62.5%). The χ^2-test of independence turned out to be highly significant suggesting that successful first contact and day of the week of the first contact attempt are not independent ($\chi^2 = 35.6$, $df = 6$, $p = 0.0001$). Wednesdays were the least successful days for first contacts in France, only leading to success in 40.9% of attempts, whereas Fridays were the most successful days with a successful first contact rate of 45.0%. A χ^2-test of independence yielded a *p*-value of 0.7188 suggesting a highly likely independence between the day of the first contact and the outcome of the first visit ($\chi^2 = 2.9$, $df = 6$, $p = 0.7188$).

On average, UK sample units who were successfully contacted at the first attempt were visited later in the day than successfully contacted units. This difference was significant ($p<0.0001$; see Table 10). On the other hand, a comparison of the groups' average first contact attempt times showed no significant differences for the German sample with non-contacted units being visited at almost the same time as successfully contacted units ($p=0.0600$). Contrary to previous findings (Campanelli et al. 1997; Durrant et al. 2011; Durrant and Steele 2009; Lipps and Benson 2005; Purdon et al. 1999; Stoop 2005, p. 160f; Vicente 2017; Wagner 2013; Wang et al. 2005; Weeks et al. 1987), successfully contacted units were visited slightly *earlier* on average than non-contacted units in France, which proved to be a significant mean difference ($p<0.0001$; see Table 10).

	United Kingdom			**Germany**			**France**		
	successful contacts %	**row n**	χ^2 ***df*** ***p*-value**	**successful contacts %**	**row n**	χ^2 ***df*** ***p*-value**	**successful contacts %**	**row n**	χ^2 ***df*** ***p*-value**
Weekday of the first visit *(n)*		5,489			8,471			4,261	
Monday	43.44	1,052	22.08 6 0.0012	70.87	1,466	35.62 6 0.0001	42.15	465	2.87 5 0.7188
Tuesday	42.34	1,162		67.01	1,449		41.41	454	
Wednesday	43.44	937		67.31	1,389		40.93	474	
Thursday	41.45	731		62.49	1,349		43.17	366	
Friday	40.90	714		63.04	1,020		45.02	422	
Saturday	49.57	579		69.83	1,551		40.96	2,080	
Sunday	52.23	314		65.18	247		0	0	

Table 9: Bivariate Analysis for Fieldwork-Related Variables. (n) Represents Valid Cases.

	United Kingdom			**Germany**			**France**		
	non-contact	**contact**	***p*-value**	**non-contact**	**contact**	***p*-value**	**non-contact**	**contact**	***p*-value**
Fieldwork variables									
First contact hour	14.29	14.58	0.0001	15.18	15.30	0.0600	15.95	15.44	0.0001
Interviewer variables									
Workload	25.01	25.41	0.1701	47.13	53.14	0.0001	27.97	27.90	0.8337
Completion rate	35.72	39.20	0.0001	26.98	27.79	0.0037	43.98	48.44	0.0001
Interviewer age	59.00	57.97	0.0233	61.46	60.74	0.1585	53.78	54.40	0.1548
Respondent Variables									
# household members	2.17	2.45	0.0001	2.41	2.64	0.0001	2.13	2.32	0.0007
Years of education	14.41	14.06	0.0281	14.56	14.18	0.0169	13.40	12.85	0.0037
Household income	5.21	5.16	0.7536	5.97	6.10	0.3361	4.94	5.03	0.5222
Respondent's age	51.19	53.46	0.0042	49.94	49.56	0.6590	51.34	53.33	0.0185

Table 10: Student's t-test Results for Group Mean Differences between Contacted and Non-contacted Units. (n) Represents Valid Cases.

Table 11 shows cross-tabulations for the outcome variable of the share of successful first contacts and the set of housing- and area-related variables. When comparing the self-assessment of UK respondents with regard to their area's urbanicity, successful first contact rates were highest where respondents said they live in a farm or in the countryside. The higher the respondent-assessed urbanicity, the lower the share of successful contacts. This finding is supported by a significant χ^2-test suggesting an association between these two variables (χ^2 = 13.1, *df* = 4, *p* = 0.0109). The highest share of successful contacts was made with respondents living in single-unit houses (48.4%) rather than multi-unit houses (34.0%; χ^2 = 62.0, *df* = 2, *p* = 0.0001). Contact rates for houses in bad compared to those in good condition were roughly equal at about 47%. However, units living in houses of 'satisfactory' condition, were only contacted in 42.2% of first contact attempts. A significant χ^2-test suggests the lack of independence between these variables (χ^2 = 9.8, *df* = 2, *p* = 0.0073). Units living in buildings with access impediments were only successfully contacted in 33.5% of first contact attempts, contrary to units living in buildings without access impediments (47.8%; χ^2 = 50.4, *df* = 1, *p* = 0.0001). The amount of litter or rubbish and graffiti or vandalism in an area as well as whether a telephone number of the sample unit was available, were independent from the outcome variable in the UK sample.

German units living in farms or in the countryside were contacted at the first attempt in 81.4% of cases. The least successful contact rate was observed for respondents assessing their living area as a 'town or small city' (68.3%). A significant χ^2-test suggests that respondent assessed urbanicity and the outcome variable are not independent from one another (χ^2 = 17.6, *df* = 4, *p* = 0.0014). Those living in single-unit houses as well as those living in 'other' houses showed a higher successful contact rate compared to those living in multi-unit houses (χ^2 = 58.2, *df* = 2, *p* = 0.0001). When there were no access impediments, units were contacted during the first attempt significantly more often than houses secured with access impediments (72.1% versus 64.1%; χ^2 = 42.6, *df* = 1, *p* = 0.0001). Units for whom a telephone number was available were more likely contacted at first attempt (71.2%) compared to those without a telephone number (65.7%; χ^2 = 21.7, *df* = 1, *p* = 0.0001). χ^2-tests did not yield significant results at the 5%-significance level for the housing condition, amount of litter or rubbish and vandalism or graffiti in the area, suggesting the independence between these variables and the outcome variable in the German sample.

The highest rate of first successful contact attempts in France was made in villages (56.7%) compared to the lowest success rate in respondent-assessed areas of big cities (40.6%; χ^2 = 25.5, *df* = 4, *p* = 0.0001). A significant χ^2-test result showed the lack of independence between the type of house a unit lives in and the outcome variable, with units living in single-unit houses being successfully contacted in 48.2% of the cases compared to units living in multi-unit houses only being successfully contacted in 34.0% of the cases (χ^2 = 96.9, *df* = 2, *p* = 0.0001). Units were significantly more likely to be contacted if they were living in houses of good or very good condition (χ^2 = 28.5, *df* = 2, *p* = 0.0001) and if these houses did not have any access impediments for the interviewer (χ^2 =70.0, *df* = 1, *p* = 0.0001). Area variables of litter and vandalism do not seem to be related to the outcome variable in the French sample.

	United Kingdom			**Germany**			**France**		
	successful contacts %	**row n**	***χ²*** ***df*** ***p***	**successful contacts %**	**row n**	***χ²*** ***df*** ***p***	**successful contacts %**	**row n**	***χ²*** ***df*** ***p***
Urbanicity *(n)*		2,165	13.07 4 0.0109		2,353	17.59 4 0.0014		2,009	25.53 4 0.0001
Farm/countryside	67.53	77		81.36	59		57.25	131	
Country village	57.76	464		76.86	713		56.73	557	
Town/small city	53.64	990		68.33	843		50.50	697	
Suburbs/outskirts	50.57	441		71.54	376		54.51	277	
Big City	48.19	193		69.61	362		40.63	347	
Type of house *(n)*		4,872	61.95 2 0.0001		7,909	58.18 2 0.0001		4,207	96.92 2 0.0001
Single unit	48.39	3,873		70.00	4,100		48.28	2,465	
Multi-unit	34.03	905		62.01	3,777		34.03	1,716	
Other	40.43	94		78.12	32		7.69	56	
Condition of building/house *(n)*		4,822	9.82 2 0.0073		7,906	2.58 2 0.2749		4,261	28.54 2 0.0001
Bad or very bad	46.75	169		67.05	349		36.89	122	
Satisfactory	42.18	1,368		64.71	1,944		32.24	611	
Good or very good	47.18	3,285		66.67	5,613		43.54	3,528	
Access impediments *(n)*		4,804	50.36 1 0.0001		7,868	42.60 1 0.0001		4,261	69.96 1 0.0001
No access impediments	47.84	4,093		72.08	2,027		47.98	2,153	
Access impediments	33.47	711		64.12	5,841		35.34	2,108	
Amount of litter in vicinity *(n)*		4,829	0.86 1 0.3536		7,911	0.37 1 0.5435		4,261	0.74 1 0.3907
None, almost none or small	45.80	4,727		66.28	7,548		41.82	4,175	
Large or very large	41.18	102		64.74	363		37.21	86	
Amount of vandalism in vicinity *(n)*		4,829	0.38 1 0.5381		7,977	0.75 1 0.3880		4,261	0.64 1 0.4234
None, almost none or small	45.67	4,808		66.28	7,743		41.89	4,201	
Large or very large	52.38	21		63.10	168		36.67	60	

Telephone number available *(n)*		1,890	1.15		8,471	21.71		0	
No	46.15	39	1	65.66	6,444	1	0.00	0	-
Yes	54.78	1,851	0.2842	71.24	2,027	0.0001	0.00	0	

Table 11: Bivariate Analysis for Housing and Area Related Variables. (n) Represents Valid Cases.

	United Kingdom			**Germany**			**France**		
	successful contacts %	**row n**	χ^2 ***df*** ***p***	**successful contacts %**	**row n**	χ^2 ***df*** ***p***	**successful contacts %**	**row n**	χ^2 ***df*** ***p***
Sex of first interviewer *(n)*		1,873	0.04		2,217	0.37		1,945	2.27
Female	55.43	985	1	73.34	874	1	50.95	1,207	1
Male	54.95	888	0.8360	72.97	1,343	0.8477	54.47	738	0.1316

Table 12: Bivariate Analysis for Interviewer-Related Variables. (n) Represents Valid Cases.

For all countries under observation, no significant test result of independence between the sex of the first contact interviewer and the outcome of the first visit was observed, suggesting that interviewer's sex and first contact success are independent (see Table 12).

Table 10 shows that in the UK, units who were successfully contacted at first contact attempt were visited by younger interviewers. Even though the difference is minor and might not be of practical relevance, it is statistically significant (59.0 years versus 58.0 years; p = 0.0233). Additionally, contacted units were visited by interviewers who have a significantly higher completion rate (39.2% versus 35.7%, p = 0.0001). Further, no significant group mean differences between contacted and uncontacted units was found for the interviewer first contact workload. In Germany, successfully contacted units at first attempt were visited by interviewers with a significantly higher first contact workload (53.1% first contact workload versus 47.1% first contact workload; p = 0.0001) as well as by interviewers who had a significantly higher completion rate (27.8% versus 27.0%; p = 0.0037). No significant group mean difference was found for the interviewer's age between the groups. Successfully contacted units in the French sample were visited by interviewers with a significantly higher completion rate (48.4% versus 44.0%; p = 0.0001). No significant group mean differences were found for interviewer workload or interviewer age.

Table 10 and Table 13 show results of the tests between the respondent-related variables and the outcome variable of successful contact. In the UK, the sex of sampled units, whether they have children at home as well as their marital status seem to be independent from the outcome variable as χ^2-test results did not reach statistical significance. However, there seems to be an association between the unit's main activity in the last seven days and the outcome variable. Amongst others, units in the UK who are in paid work get successfully contacted in 49.4% of the cases, whereas sick units get contacted in 63.9% of the cases ($\chi^2 = 24.3$, $df = 8$, $p = 0.0001$). Successfully contacted units were also significantly older (51.2 versus 53.5; $p = 0.0042$), lived in households with significantly more members (on average 2.5 versus 2.2; $p = 0.0001$), and had a slightly but significantly shorter duration of education (14.0 years versus 14.4; $p = 0.0281$). No significant group mean difference was found for the household income. In the German sample, the investigated categorical characteristics of respondents seem to be independent from the outcome variable since no χ^2-tests reached statistical significance. However, contacted units lived in significantly larger households (mean 2.6 members versus 2.4 members; $p = 0.0001$) and on average underwent a slightly but significantly shorter duration of education (14.2 years versus 14.6 years; $p = 0.0169$). No significant group mean differences were found for the household income or the respondent's age. French respondents being in paid work (48.1%) or education (46.3%) during the last seven days, were contacted significantly less often than unemployed (57.3%) units or househusband and housewives (59.7%; $\chi^2 = 14.1$, $df = 6$, $p = 0.0285$). Successfully contacted units were significantly older ($p = 0.0185$), lived in larger households (2.3 members versus 2.1 members; $p = 0.0007$) and had significantly fewer years of education (12.9 years versus 13.4 years; $p = 0.0037$). No significant group mean difference was found for the household income.

	United Kingdom			**Germany**			**France**		
	successful contacts %	**row n**	χ^2 ***df*** ***p***	**successful contacts %**	**row n**	χ^2 ***df*** ***p***	**successful contacts %**	**row n**	χ^2 ***df*** ***p***
Respondent's sex *(n)*		2,165	0.65		2,353	0.36		2,010	0.59
Female	53.12	1,186	1	71.38	1,146	1	52.32	1,097	1
Male	54.85	979	0.4210	72.49	1,207	0.5473	50.60	913	0.4418
Children at home *(n)*		1,204	0.10		1,058	0.10		1,101	1.02
No	56.81	649	1	72.33	571	1	55.68	607	1
Yes	57.74	555	0.7460	73.20	487	0.7527	52.63	494	0.3119
Respondent's marital status *(n)*		1,225			1,107			1,156	
Married	50.69	144	4.51	77.36	53	2.22	36.17	17	5.30
Separated/divorced	51.36	257	3	73.30	191	3	48.25	138	3
Widowed	55.02	209	0.2110	69.54	151	0.5279	53.52	114	0.1509
None of these	46.99	615		69.66	712		47.38	289	
Main activity in the last 7 days *(n)*		2,163			2,347			2,008	
Paid work	49.38	1,134		70.37	1,198		48.11	954	
Education	54.67	75	24.25	73.01	226	6.31	46.28	121	14.10
Unemployment	55.74	61	6	76.67	60	6	57.27	110	6
Sick	63.83	94	0.0001	81.13	53	0.3893	54.93	71	0.0285
Retired	58.96	614		72.06	562		55.18	685	
Housework	61.05	172		73.89	203		59.65	57	
Other	38.46	13		80.00	45		70.00	10	

Table 13: Bivariate Analysis for Respondent-Related Variables. (n) Represents Valid Cases.

Figure 8 to Figure 13 show correlation matrices between the independent variables. The fill colour and shading represent the effect size and direction: red indicates negative Pearson's product-moment *r*-correlations or φ-coefficients, while blue indicates positive *r*-correlations or φ-coefficients. Correlations that are non-significant at a 5% significance level are blanked out. Results for categorical variables are depicted in Figure 8 to Figure 10, whereas results for continuous variables are depicted in Figure 11 to Figure 13.

Most correlations for categorical variables are non-significant or have a small effect size in the area between -0.3 and +0.3 for all countries under examination. Most of the strongest negative correlations are logical exclusions, for example being male is negatively correlated with being female, or having children at home is negatively correlated with not having children at home. These are included for sake of completeness but do not add any empirical value to the investigation. Some positive and medium-to-strong correlations include being retired or widowed and not having children at home, which is similar across countries. Buildings *without* access impediments are positively correlated with single-unit dwellings, while buildings *with* access impediments are positively correlated to multi-unit houses. This effect is stronger in the UK and France than in the German sample.

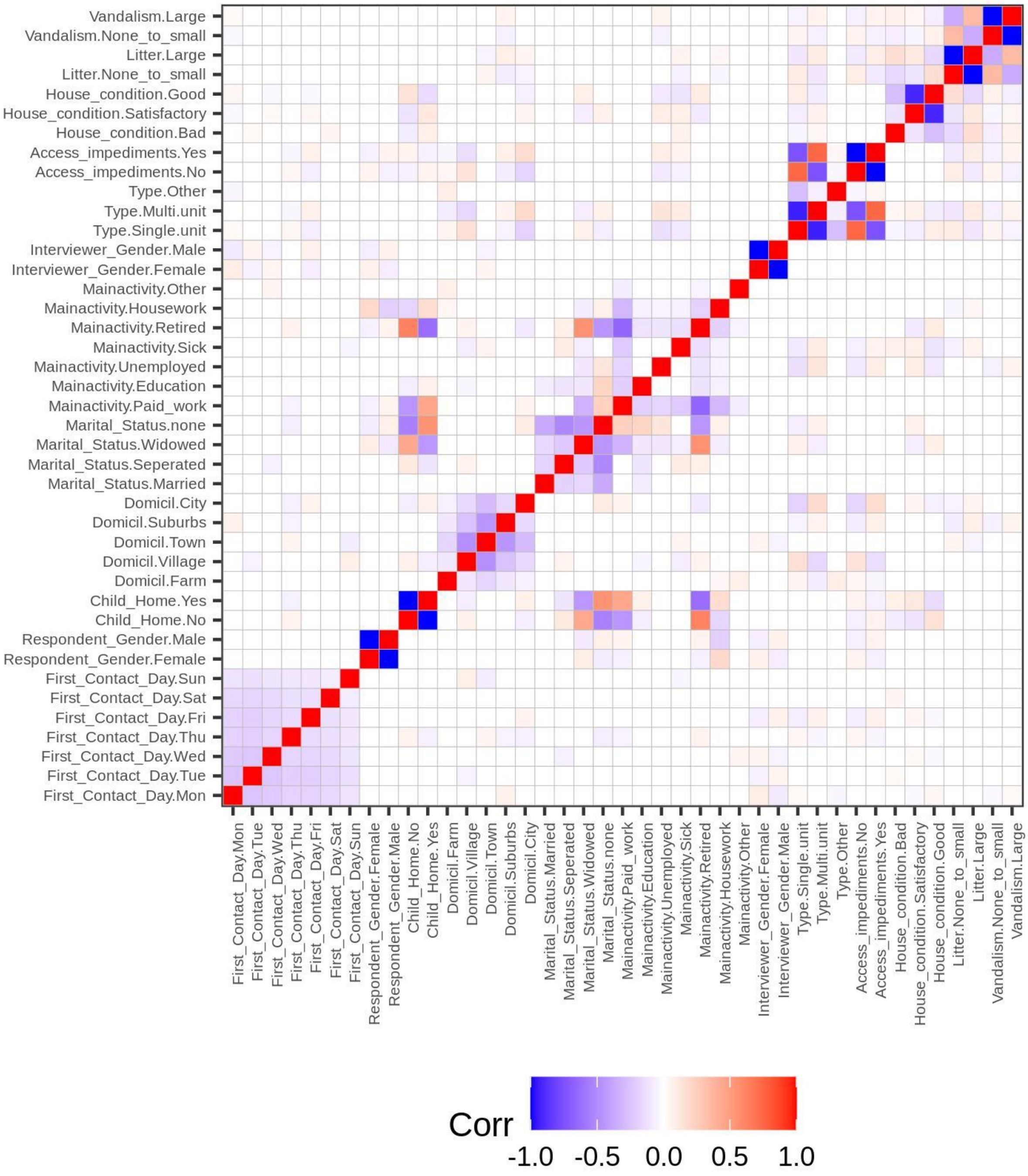

Figure 8: Correlation Matrix for Independent Categorical Variables in the UK Sample

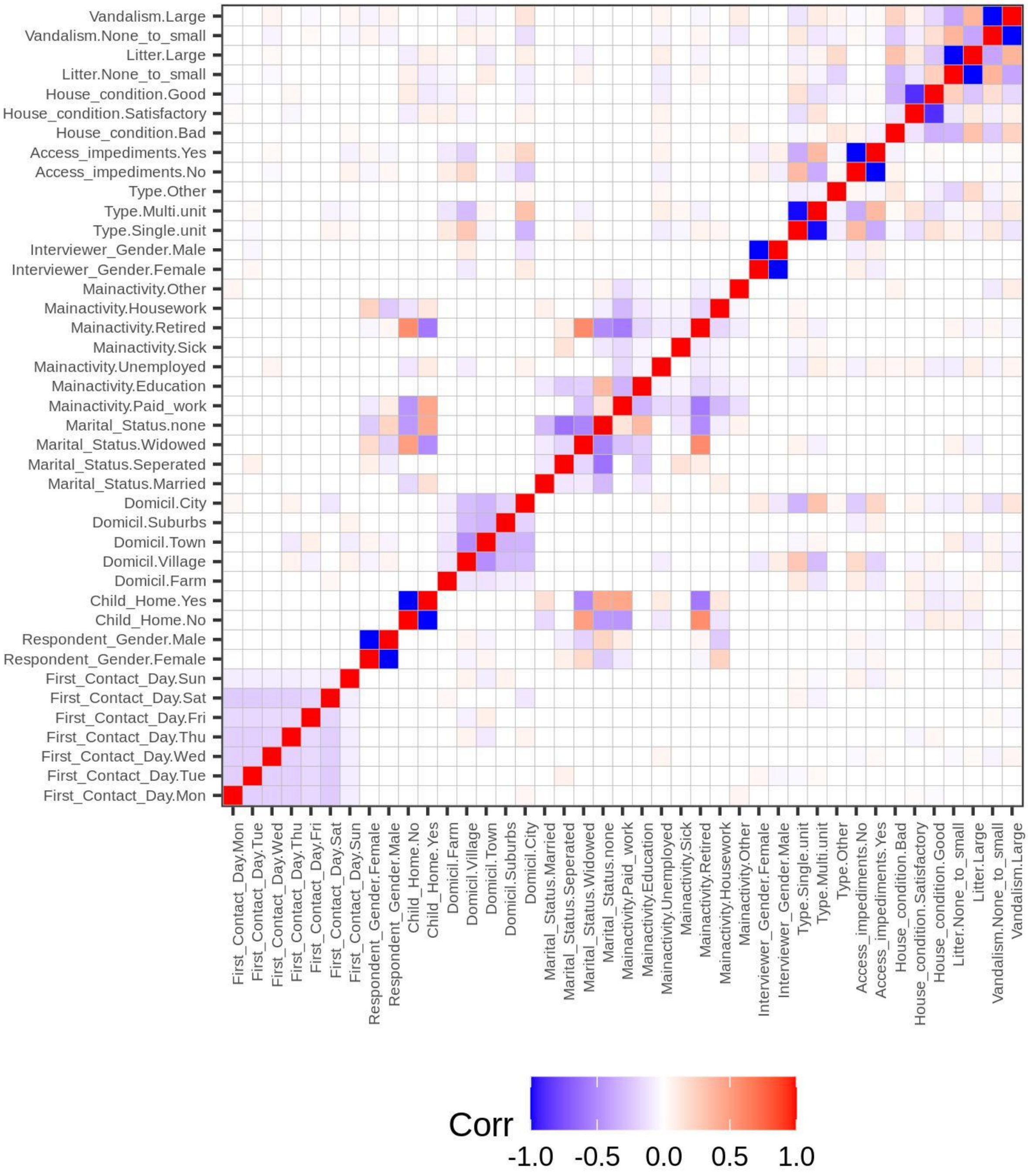

Figure 9: Correlation Matrix for Independent Categorical Variables in the German Sample

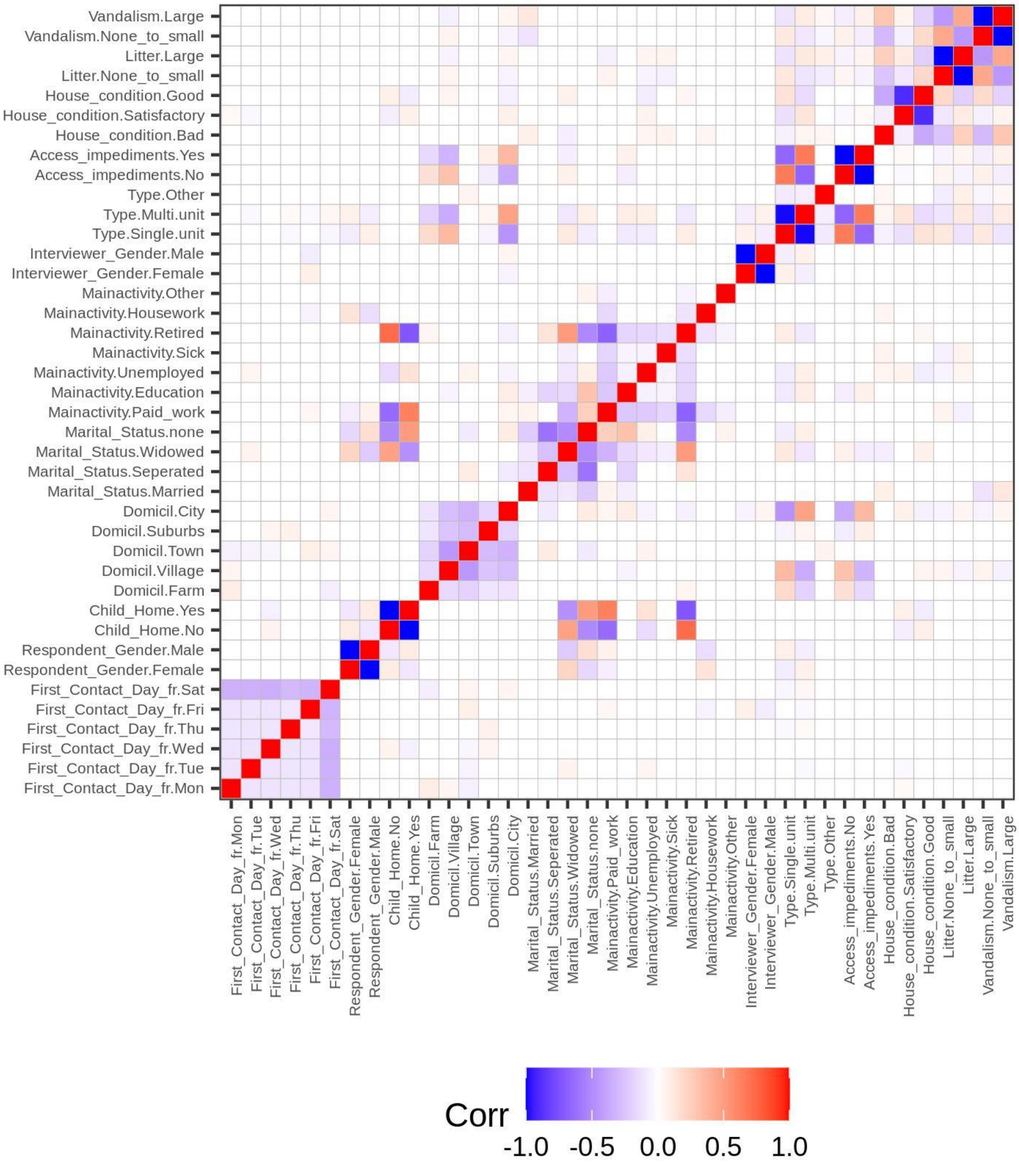

Figure 10: Correlation Matrix for Independent Categorical Variables in the French Sample

Pearson's product-moment correlations *r*-coefficients for the continuous variables are shown in Figure 11 to Figure 13. Across all countries, strong negative correlations can be found for age effects like respondent age and number of household members, respondent age and years spent in education or respondent age and household income. Strongest positive correlations can be found between household income and household members along with household income and years of education. Significant negative correlations can be found between interviewer workload and interviewer age in the UK and Germany, while this marks a significant positive correlation in France. While no significant correlation was found between the completion rate and interviewer age in France, a significant negative correlation between these variables was found in Germany and a significant positive correlation was found in the UK. The completion rate was significantly negatively correlated to workload in the UK, while in Germany and France a significant positive correlation between these variables was found. In Germany and France, a significant negative correlation was found between first contact hour and interviewer age.

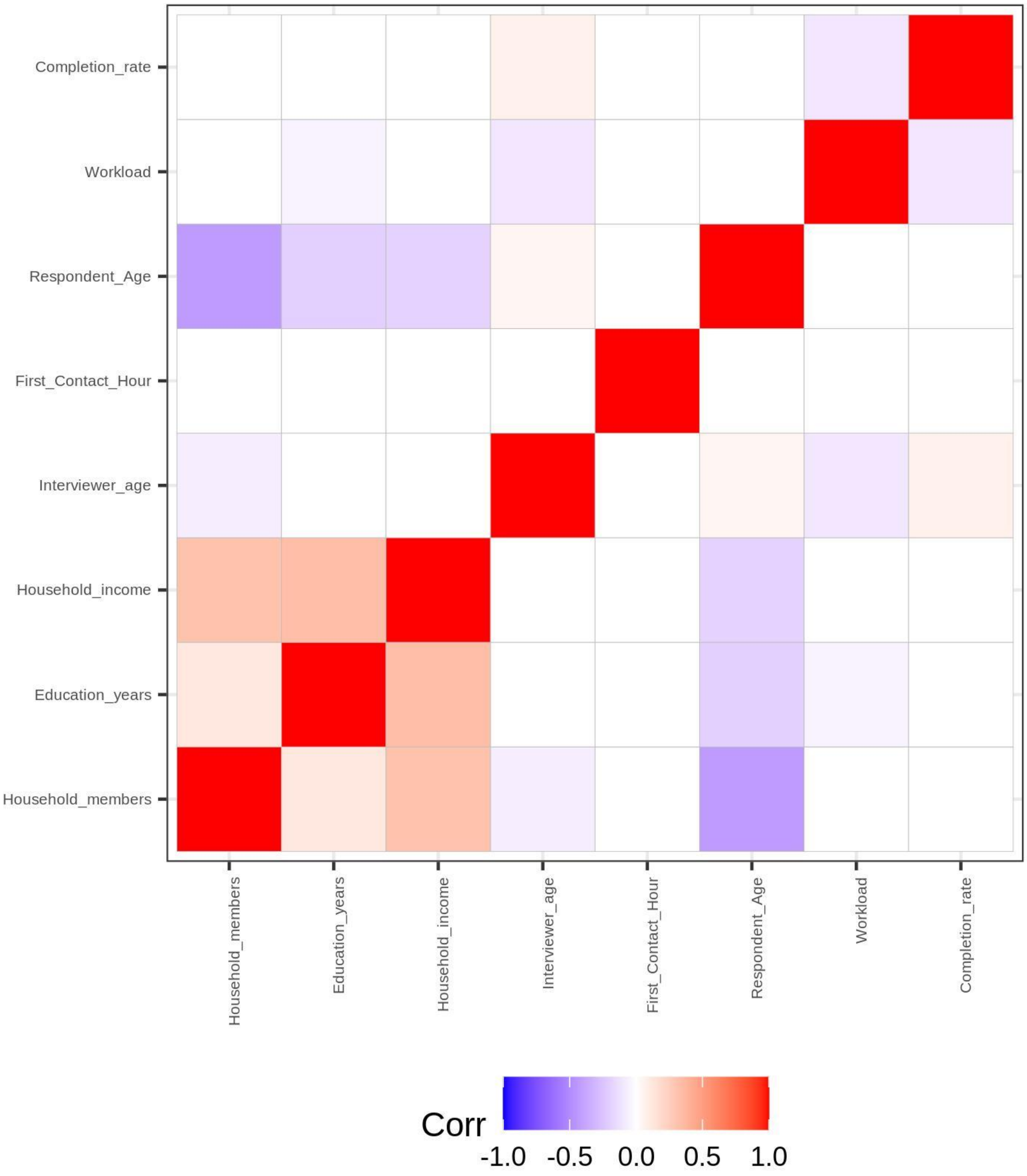

Figure 11: Correlation Matrix for Independent Numerical Variables in the UK Sample

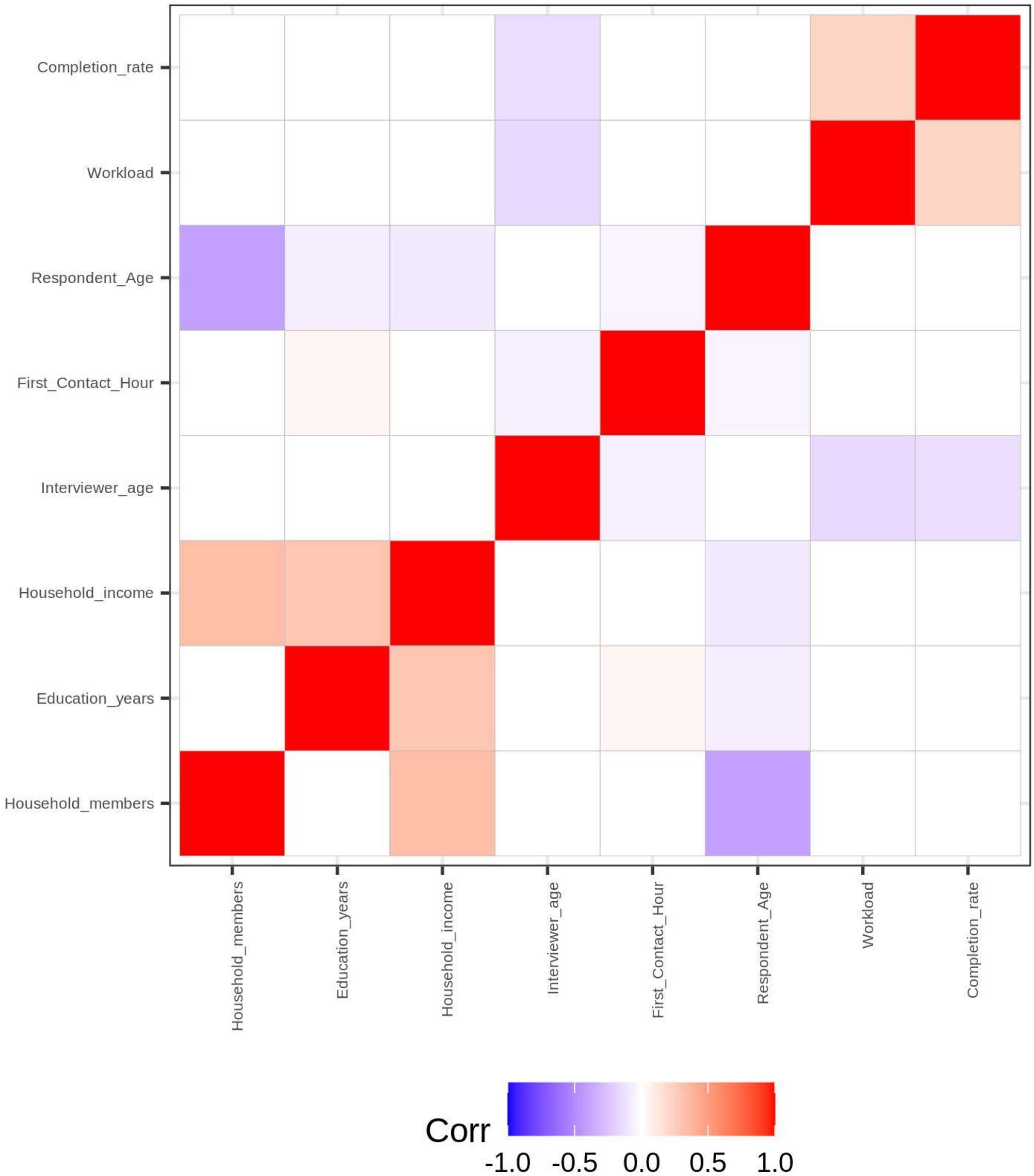

Figure 12: Correlation Matrix for Independent Numerical Variables in the German Sample

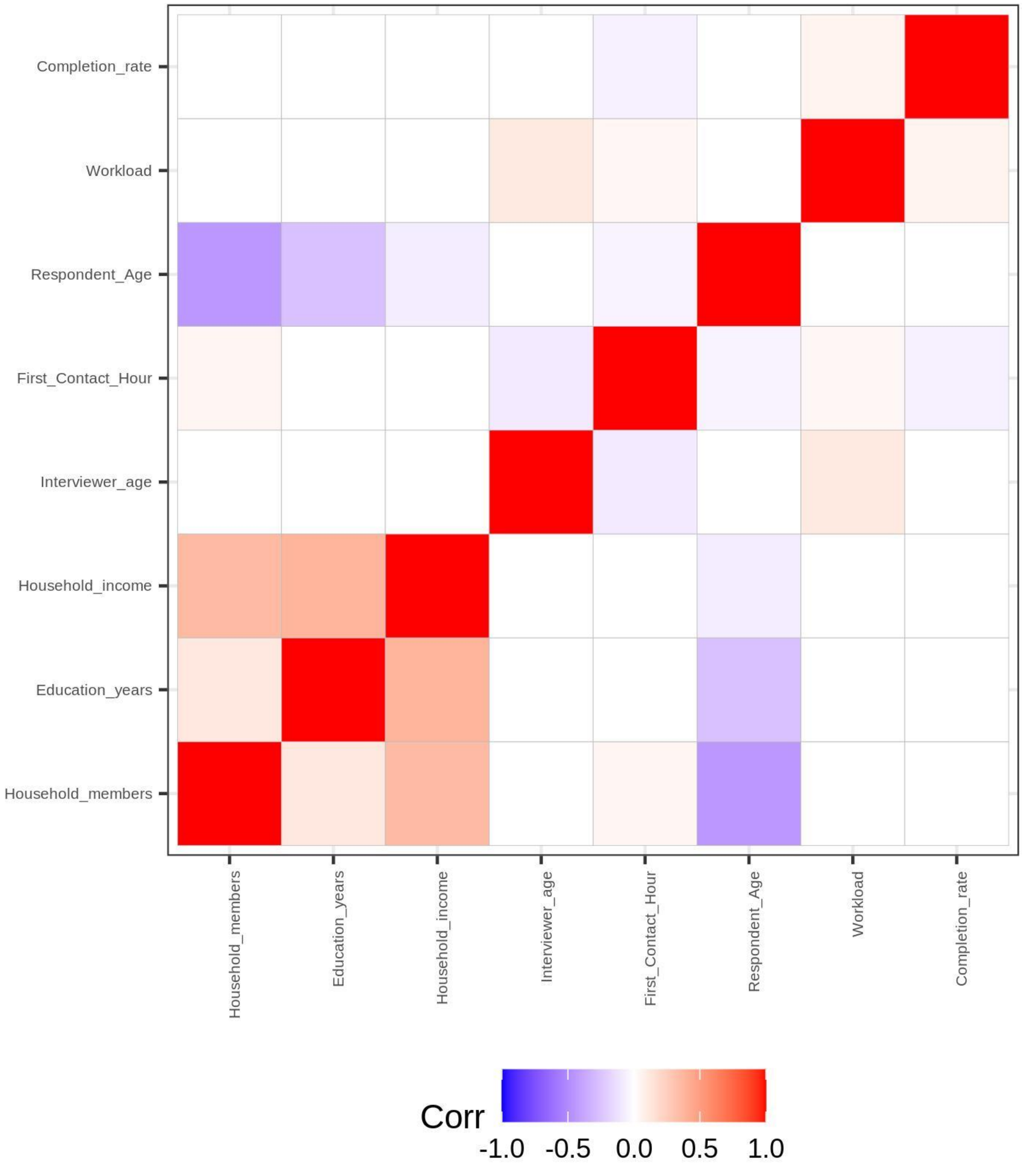

Figure 13: Correlation Matrix for Independent Numerical Variables in the French Sample

4.3. Discussion & Conclusion

The analyses of correlates of contact in ESS Round 9 produced several interesting findings. It was striking that while the German sample had a similar number of participants (2,358) compared to the UK (2,204), the German sample had a much higher number of nonrespondents (6,337 in Germany versus 3,646 in the UK), making the German sample much less efficient. Further, data selection processes showed that in the UK roughly 6.2% of the gross sample needed to be removed due to exclusion criteria. Of all countries under observation the UK sample had the most units with missing contact information sheets (36). Additionally, while the variable for the outcome of the first visit was never missing in the German and the French sample, it was missing in 29 cases of the UK sample. Invalid addresses were a common problem in all countries, but most invalid addresses were recorded in the UK sample. These results suggest that there is room for improvement in the UK fieldwork processes. Contact information sheets should be more reliably filled in by the interviewers and included in the datasets as they carry important information for researchers and fieldwork agencies alike. Additionally, the outcome of the first visit can yield important information for further contact attempts, thus interviewers should be encouraged to diligently report this outcome. Lastly, it seems that the quality of the UK's sampling frame would benefit from improvement since it features a larger number of incorrect addresses compared to the frames of the other two countries.

One of the most striking findings is that 9% of first contact attempts in France led to a completed or partial interview compared to only 1% completed or partial interviews in

Germany and 5% in the UK. While these proportions were low across all countries, there were notable percentage point differences. On the other hand, the French sample yielded the highest number of unsuccessful first contact attempts (58%). This shows that while the fieldwork processes in France might carry the largest improvement potential with regards to establishing first contact attempt success, the interviewers in France might have better strategies to convince units to participate once they established a contact. Another explanation might be that the French population is more amenable to surveys. In contrast, the German sample suffered the least from unsuccessful first contacts (33.0%), but also had the smallest proportion of completed or partial interviews after the first contact attempt was successful (1.2%). While establishing contact seems to be easier in Germany than in France, the former seems to lack the efficiency of the latter in converting contacts into participants during the first visit. Fieldwork efficiency in the UK is situated between the fieldwork efficiency of the other two countries in terms of both contact success and partially or fully completed interviews during the first contact. While in the UK there were more completed and partial interviews at first contact attempt compared to Germany (4.8%), the UK also suffered almost as much from unsuccessful contacts as the French sample (56.3%).

The analyses revealed large differences in contact procedures across countries in terms of contact days and times. The most striking finding was found in the French sample where almost half of first contact attempts were made on Saturdays and none on Sundays, while contacts for the remaining half of the sample were spread out over one of the five remaining weekdays. In the UK, there were contacts made on Sundays, although they were less prominent compared to other days. Furthermore, more than 57% of contacts happened during the first three weekdays. Only the German sample showed an even distribution of first contact attempts throughout the week – except for Sunday (with slightly fewer contacts), with roughly 16% of the sample contacted each day from Monday to Saturday. The analysis suggests an association between contact success and the day of the week of the first visit in the United Kingdom and Germany but not in France.

It is striking that first contact attempts in the UK were less prominent after around 4pm compared to the German sample. Most first contact attempts in the UK happened between around 12pm and around 4pm regardless of the day of the week. Fieldwork times in Germany and France were more similar to one another. Even though there were overall fewer contacts on weekdays in France, the units were often contacted after 5pm on weekdays and during the daytime on Saturdays. Similar patterns were found in the German sample, where contact attempts were relatively most likely between around 4pm and around 6pm on weekdays and during the daytime on Saturdays. Besides these univariate findings, when looking at the bivariate group mean comparisons of contact success between contacted and non-contacted units, the analysis shows that successfully contacted units were visited at later hours of the day than uncontacted units in the UK and German samples. In France, however, contacted units tended to be visited earlier in the day than non-contacted units. The answer to Research Question 1 (*'Is a successful contact associated with the day of the week and time of the day?'*) is thus country specific: In both the UK and Germany a successful first contact attempt is associated with both the day of the week and time of the day. In France, only an association between the time of the day was found but not day of the week. Interestingly, the direction of the association in France points in the opposite direction of the findings for the UK and Germany.

Generally, fieldwork procedures seem to follow the literature's recommendations of preferable contact times on weekday afternoons and evenings (Campanelli et al. 1997; Durrant et al. 2011; Durrant and Steele 2009; Hox et al. 2006; Lipps and Benson 2005; Purdon et al. 1999, p. 160; Stoop 2005; Vicente 2017; Wagner 2013; Wang et al. 2005; Weeks et al. 1987). Although on average it appears that the UK fieldwork follows a different strategy, since units are visited slightly earlier in the day on average than in the other countries, it remains questionable whether this finding can be attributed to different fieldwork strategies or not. Yet, it might be worthwhile for the UK fieldwork efficiency if interviewers were convinced to conduct their first contact attempts slightly later in the day to match, for example, the average first contact attempt hour of the day of the German sample. The high contact success rates on Saturdays in the French sample are interesting, since previous research is not clear about whether weekend contact attempts are more successful than weekday contact attempts (Groves and Couper 1998; Lipps and Benson 2005; Purdon et al. 1999; Stoop 2005; Vicente 2017) or not (Hox et al. 2006; Weeks et al. 1987). The high share of successful first contacts on Saturdays could also explain the high share of 'completed interview' outcomes at the first call, as units who are contacted on Saturdays might be more willing to participate in a survey compared to other days of the week for various reasons. For example, they might be less engaged with work or spend more free time at home if they do not have to work on Saturdays and thus be available for an interview. However, these ideas remain unexplored in this analysis and need further research. Although the fieldwork processes in France seem to be already largely in line with the recommendations in the literature for establishing contact, it is notable that – in contrast to both the literature and findings here for Germany and the UK – non-contacted units were visited at later hours of the day on average, whereas in the other countries, non-contacted units were visited at earlier hours of the day on average. Again, this result might be confounded with the high number of visits on Saturdays, which are less successful the later in the day they take place. Fieldwork procedures in Germany are more closely in line with the recommendations from the literature. This might be the reason for relatively high first contact success rates compared to the UK's and France's sample data.

To answer Research Question 2 (‘*With which area or household characteristics is first contact attempt success associated?’*) the analysis shed light on various variables that relate to the characteristics of the area or household a unit lives in. Findings from previous research suggested a lower contact success in more urban areas and these findings are supported by the analysis. In fact, the analysis showed that there is an association between the degree of urbanicity and contact success in all countries. The pattern is most obvious in the UK sample, which shows that as the degree of urbanicity increases, the share of successful contacts decreases. Similar patterns can be observed also in the German and French samples, but while these relationships reach statistical significance, their pattern is less obvious compared to the UK. The most successful contacts in Germany and France were made in the countryside or villages, while in towns or small cities fewer successful first contacts were made. Interestingly, in areas which respondents described as 'suburbs or outskirts of a big city', the share of successful first contacts was much higher compared to the areas described as 'town or small city' or 'big city'. Furthermore, the share of successful first contacts in suburbs or outskirts was almost as high as in 'country villages'. This finding adds to the body of literature that conducting surveys in areas with a higher degree of urbanicity is more complicated (Campanelli et al. 1997; Durrant et al. 2011; Groves and Couper 1998; Hox et al. 2006; Luiten and Schouten 2013; Robinson and Godbey 1997; Stoop 2005). Overall, the findings unanimously point in the same direction in all countries with lower contact success in more densely populated areas. However, previous research had found country differences in the directions: Blom (2012), who investigated contact success in general and not only the first contact, found no effect between the degree of urbanicity and contact probability in the UK, Belgium and Greece. While in Finland and Ireland rural areas were positively associated with contactability, the association in Portugal was negative. While these opposing findings are noteworthy, they might be due to a different operationalisation of 'urbanicity', since Blom did not use the respondent's description of the house as a proxy for the degree of urbanicity like in this analysis, but rather defined urbanicity as the 'percentage of farms and single housing units in the first assignment of the interviewer making the first contact attempt to a sample unit' (Blom 2012, p. 220).

Single-unit houses were found to have a much higher share of contact success compared to multi-unit houses in all three countries – a finding which is also in line with the literature (Campanelli et al. 1997; Stoop 2005). This adds to answering Research Question 2 as contact success is associated with the type of house a potential respondent lives in. Since correlations showed that multi-unit houses are positively correlated with the degree of urbanicity, this further contributes to the finding of a lower first contact attempt success in more urban areas. Findings from previous research showed that contact propensity is reduced in areas with higher crime rates (Groves and Couper 1998). Unfortunately, data on an area's crime-rate was not available for this analysis. However, in the absence of such data, an assumption was made that deprived areas (indicated by signs of decay like a larger amount of litter and rubbish as well as a higher amount of vandalism and graffiti) serves as a proxy for higher crime-rate areas. Yet, the results do not suggest an association between first contact attempt success and whether the units live in deprived areas or not in any country. However, this this finding might change with increases in the sample size for areas, which show signs of decay, since cell sizes were rather small. From a fieldwork perspective, finding no differences between areas, which show signs of decay and wealthier areas is reassuring since it suggests that interviewers might not apply different strategies between these areas. This though, however, needs to be confirmed in more advanced analysis and under control of more relevant independent variables.

Units in houses of either very good or very poor physical conditions had significantly higher contact rates than those living in satisfactory housing conditions in the UK and France, and the results suggest an association between first contact attempt success and the physical condition of a unit's building or house in those countries. The suggested direction of this finding is interesting because previous research predominantly found a negative relationship between contact success and houses in a bad physical condition (Blom 2012; Durrant, D'Arrigo, and Steele 2011; Durrant and Steele 2009; Hox, Blohm, and Koch 2006; Lipps and Benson 2005;

Stoop 2005, p. 175). Interpretations of this finding are difficult, but the physical condition of a house might be related to socio-economic factors like employment status, type of occupation or income. One possible explanation is that units living in worse-off houses also spend more time at home, because of irregular working times or even unemployment, and are thus more likely to be contacted. These presumptions cannot be investigated here but might be of interest for future research.

A clear pattern can be also found for houses with access impediments for which first contact success rates were significantly lower in all countries, which suggests an association between first contact attempt success and whether or not a house had access impediments. This finding is largely in line with the literature (Blakely and Snyder 1997; Cunningham et al. 2005; Durrant et al. 2011; Groves and Couper 1998, p. 88; Hox et al. 2006; Lipps and Benson 2005; Wang et al. 2005), with only a few investigations pointing to varying, none or even positive influences for some countries (Blom 2012; Durrant and Steele 2009). Following this finding in the ESS, fieldwork conducted in areas with higher shares of houses with access impediments must be expected to face increased data collection costs due to a higher possibility of necessary revisits. The analysis revealed a positive correlation between access impediments and multi-unit houses, indicating that areas with a particularly high share of multi-unit houses might be prone to less first attempt contact success. This finding contributes to the understanding of higher non-contactability in more urban areas, which was also shown by the positive correlation between multi-unit houses and degree of urbanicity.

Next, previous research showed mixed results on initial calls – either concluding that establishing a phone call before visits increases contact success rates (Lipps and Benson 2005; Schnell 1997, p. 220) – or that not having a unit's telephone number is a practical impediment rather than an indicator for non-contactability (Stoop 2005, p. 169) or further, finding varying results per country in other cross-national studies (Blom 2012). This investigation shows that having a telephone number was not recorded in the French sample at all, while in the UK sample the variable is only available for respondents and only the German sample includes the variable for both respondents as well as nonrespondents. Whether or not contact success is associated with the availability/absence of a telephone number was thus not testable in the French sample. In the UK no association was found while the analysis points towards an association in the German data. Since this variable might be derived from the sampling frame or is produced in the sampling stage, it might simply not be available in all countries because of the different frames that are used. Since this variable seems not to be harmonised across countries i.e., it remains unverifiable for the French sample as well as it includes rather low cell sizes in the UK sample, it will not be used in later analyses of this thesis.

Summing up the investigations for Research Question 2 it shows that some parallels between the countries can be found: While the analysis showed that there is an association between degree of urbanicity, the type of house or whether or not access impediments are present and the first contact attempt success in all countries, no association was found for whether or not the area showed signs of decay and contact success. However, some findings remain context specific as an association between the physical condition of a house and contact success can be found in the UK and France but not Germany. Similarly, an association between the presence/absence of a telephone number and first contact attempt success can be found in Germany but not in the UK.

The analyses of interviewer related variables to answer Research Question 3 (*'With which interviewer characteristics is first contact attempt success associated?'*) showed interesting associations between the countries' fieldwork operations and interviewer associated characteristics. When comparing the number of unique first contact interviewer identification numbers, it is striking that a lot more interviewers were conducting the first contact attempts in the UK compared to the other countries. This becomes particularly obvious when taking the gross sample sizes into account. The higher number of interviewers in the UK leads to the lowest first contact workload in the UK compared to all countries, while in the German sample more than twice the amount of first contact workload needs to be done per interviewer. Even though previous research showed a negative association between increasing workload and contact success (Botman and Thornberry 1992; O'Muircheartaigh and Campanelli 1999), the highest first contact success was actually achieved in Germany, despite having the highest interviewer workload compared to the other countries under examination. This finding is supported by Blom (2012), who predominantly found no negative association between workload and contact propensity and argues that exceptionally good interviewers are allocated a higher number of interviews and that their ability to establish contact outweighs the workload burden. Interestingly, within the German sample, those units who were successfully contacted, were contacted by interviewers with a significantly higher first contact workload compared to those interviewers who unsuccessfully contacted units – this supports Blom's (2012) findings. Although results for both the UK and France did not reach statistical significance, they are interesting as they show a higher interviewer first contact workload for contacted units in the UK, while in the French sample the interviewer workload appears to be almost the same for contacted and non-contacted units. The contradicting findings in the literature and in this analysis, however, could be the result of a different operationalisation of 'interviewer workload'.

To approximate an interviewer's success, a completion rate was used, defined as the number of all interviews the interviewer completed divided by the total number of all first contacts the interviewer needed to do (first contact workload) multiplied by 100. Interviewers who had a higher rate of completed interviews relative to their total first contact workload were deemed to be more successful interviewers. The analysis showed significant group mean differences between contacted and non-contacted units in all countries. This finding shows that units who were contacted, tended to be visited by 'more successful' interviewers, i.e., those who had a higher share of completed interviews relative to their total first contact workload. Although the concepts of measuring a 'successful' interviewer varies largely in the literature, the findings contribute to the evidence of other studies indicating that interviewers who are more successful in completing interviews relative to their workload are also more successful in establishing contact, even though country differences seem to exist (Blom 2012; Durrant, D'Arrigo, and Steele 2011; Durrant and Steele 2009; O'Muircheartaigh and Campanelli 1999).

The bivariate analyses between an interviewer's sex and their share of successful first contact attempts did not yield any significant results in any country, which supports previous findings (Blom 2012) and suggests that first attempt contact success is not associated with the interviewer's sex. Further, a significant difference in the interviewer's age between those who were contacted and those who were not, could only be found in the UK sample, suggesting that contacted units were visited by significantly younger interviewers. Despite the fact that this difference proved to be significant, the difference was so small that the practical relevance for it remains questionable. In fact, findings in the German sample point in the same direction but do not reach statistical significance, while contacted units in France, on the other hand, were visited by older interviewers. The direction of the significant age difference in the UK sample contradicts findings from previous literature, which showed higher contact rates for older instead of younger interviewers (Durrant et al. 2011; Hox et al. 2006).

To summarise the analysis for Research Question 3 it can be noted that identical findings were found for two of the investigated variables: For all countries, the analysis suggests an association between the completion rate and first contact attempt success, while no association seems to exist between the interviewer's sex and first contact attempt success. The results for the other two variables are country specific: a statistically significant association between workload and first contact attempt success can only be found in Germany, while the age of an interviewer only appears to be associated with first contact attempt success in the UK.

Lastly, analyses of the respondent level variables to answer Research Question 4 (*'With which potential respondent characteristics is first contact attempt success associated?'*) marked the independence between a respondent's sex and their successful contact in all countries. This finding contradicts results from previous research, which found higher overall contact success rates for women (Groves and Couper 1998, p. 136; Stoop et al. 2010, p. 14). While previous research related the higher contact success of women to labour-market differences i.e., a higher involvement of women in at-home care duties, these differences might have shrunk in the recent years and led to the disappearance of this dependence in this study. Another possible explanation for this deviation is that contact success in the first attempt has only rarely been the primary dependent variable of analyses. Instead, most research focused on the overall contact success or even on later visits, while in this dissertation, the first contact attempt success is the explicit focus. Previous research has also found a significant relationship between a unit's contactability and whether or not they have children at home (Durrant et al. 2011; Durrant and Steele 2009; Groves and Couper 1998, p. 91; Luiten and Schouten 2013; Lynn and Clarke 2002; Stoop 2005, p. 174). This study, however, found a likely independence between these variables, suggesting that having children at home is not associated with first contact attempt success in any country. The theorised mechanism of the association between these variables appears to be plausible and the empirical evidence of previous research is substantial, thus it is surprising that this investigation does not find support for this idea. An explanation might be that the operationalisation for having children at home (see Chapter 3.3) which was used in this study does not grasp the actual phenomenon.

In addition, a unit's contactability does not seem to be associated with their marital status in any country. Contacted households are slightly larger on average than non-contacted households in all countries. This finding is not only supported by the literature (Stoop 2005), it also suggests that it is the sheer number of persons living in a household that might increase the likelihood of contact instead of specific characteristics that are unique to units of a specific marital status. Similar to previous research (Stoop 2005, p. 180), a unit's contactability seems to be associated with their employment status in the UK and France, while this relationship does not reach statistical significance in the German sample albeit descriptive differences are in a similar direction. The findings from a Dutch study of lower contact success for higher educated units (Stoop 2005, p. 66) find support in the analysis for all three countries since units who were contacted at first attempt had spent significantly less years in education on average. The average household income decile did not vary significantly between contacted and non-contacted units in any of the countries under investigation. Units who were contacted were significantly older than the non-contacted ones in the UK and France but not in the German sample, where no significant group mean difference was found. While the findings in the UK and French samples are supported by the literature (Luiten and Schouten 2013; Vicente 2017; Weidman 2010), the findings for the German data remain distinctive. Further research is needed to address these findings in more depth and provide clarity to what the cross-country differences are attributed to.

Interestingly, the answers for Research Question 4 are less country specific but instead show parallels between the countries in six of the eight variables. There seems to be no association between a respondent's sex, whether children live at home or not, the marital status or the household income and first contact attempt success in any country. However, associations exist between the number of household members or the years spent in education and first contact attempt success in all countries. The employment status as well as the unit's age is only associated with contact success in the UK and France, but not in Germany.

For clarification Table 14 summarises the overall results of the analysis and answers to the research questions for each country. If an association was found the cell is colour-coded in dark grey while cells colour-coded in light grey indicate that no association was found. One single colour-code per line shows that the results were unanimously across countries. This is true in 12 cases, which underlines the importance of contextualisation when it comes to investigating first contact attempt success.

Research Question	Is first contact attempt associated with…	UK	Germany	France
1	… the day of the week?	Yes	Yes	No
	… the hour of the day?	Yes	Yes	No
2	… the degree of urbanicity?	Yes	Yes	Yes
	… the type of house a respondent lives in?	Yes	Yes	Yes
	… signs of decay in an area like the amount of litter/rubbish and vandalism/graffiti?	No	No	No
	… the physical condition of a unit's building or house?	Yes	No	Yes
	… whether or not a house has access impediments?	Yes	Yes	Yes
	… whether or not a telephone number is available?	No	Yes	Not verifiable
3	… the interviewer workload?	No	Yes	No
	… the completion rate?	Yes	Yes	Yes
	… an interviewer's sex?	No	No	No
	… the age of an interviewer?	Yes	No	No
4	… a respondent's sex?	No	No	No
	… whether or not a respondent has children at home?	No	No	No
	… a respondent's marital status?	No	No	No
	… the number of household members?	Yes	Yes	Yes
	… a unit's employment status?	Yes	No	Yes
	… the years a unit spent in education?	Yes	Yes	Yes
	… the household income?	No	No	No
	… the unit's age?	Yes	No	Yes

Table 14: Summary of Answers to Research Questions.

This first chapter of the second part of this thesis has investigated the factors associated with first contact success, using data for three countries from the latest European Social Survey Round 9. The investigation drew on literature on 'correlates of contact' and derived four high-level core research questions to establish which variables are associated with first contact attempt success in the ESS Round 9. Above all the analysis revealed large between-country differences for the investigated variables and their relationship to first contact success. The analysis supports multiple findings from the literature and further endorses the evidence that specific fieldwork processes and circumstances are not only highly confounded with one another, but also vary across countries. While some implications can surely be derived from the analysis it is important to note that this analysis focused on investigating correlations and not causations. Additionally, only univariate, and bivariate analyses were conducted which means that results might change when applying more complex statistical analyses and controlling for multiple variables simultaneously. For example, it was shown that both time of the day and day of the week of the first contact not only have their own bivariate association but also a conditional association on first contact attempt success. This investigation serves as a starting point to explore the relationship between specific fieldwork practices, area and household, interviewer, and respondent characteristics and contact success. The findings from this analysis will be extended in the next chapter to leverage machine learning models to predict first contact attempt success using ESS Round 9 data.

5. Can We Forecast Who Gets Contacted?

Predicting First Contact Attempt Success in the ESS Round 9

Chapter 4 already gave some insights into the mechanisms that influence first contact attempt success and showed the correlates of contact, making use of the latest available ESS data. Logically, these insights are based on data from the past and are a valuable resource for contributing to a better understanding of these associations. However, fieldwork institutes would benefit even more if it were possible for them to know the outcome of a contact attempt in advance to plan costs and/or logistics more precisely.

Instead of investigating the associations between independent variables and the first contact attempt success, this chapter focuses on the *prediction* of contact success at the first attempt instead. To accomplish this, the variables, which were derived in the literature review and presented in Chapter 4, will be fed to various algorithms. The predictions will then be compared based on their quality of correctly predicting successful contacts. Within this competition, one focus will be on investigating whether a logistic regression is able to outperform more complex machine learning algorithms with regard to the predictive performance.

While previous analyses in Chapter 4 were structured by the context of a variable, for example whether a variable was area related or interviewer related, the following analyses are structured by allocating the variables to four different input datasets per country that are then fed to the various machine learning algorithms. These input datasets are referred to as ‘models’ from here onwards. The models vary considerably in the number of variables they contain. The reasoning behind this approach is twofold: on the one hand, machine learning algorithms work best (or only) with complete cases and apply listwise deletion if necessary. To reduce the risk of incomplete cases which increases with a higher number of variables, the number of variables gets reduced from Model 1 to 4. On the other hand, by prioritising the exclusion of variables that are only available for respondents, the models rely less and less on substantive survey interview variables from Model 1 to Model 4. Since between the models the number of variables decreases steadily, the risk for a sharp decrease in sample size due to listwise deletion gets reduced and the number of observations per model can increase[13]. Thus, the technical reason behind building these four models is to increase the available sample size after listwise deletion by decreasing the necessary input variables. At the same time, the theoretical reason is to shrink the models so that the smaller a model becomes, the more it bases the predictions on paradata only and the less it relies on substantive survey information, which is only available for survey participants. These smaller models are also more applicable to real-world scenarios, in which data on a unit, who was not yet contacted, is simply not available, while paradata, even only as a proxy, might be at hand. In other words, this approach maximises the number of observations and minimises the information that is only available for respondents.

[13] In fact, the caret package features methods to replace missing values for example through *k*-nearest neighbour imputation. While Proof-of-Concept analyses that used these imputation methods yielded much better results, the approach was not pursued. In typical data science use cases the researcher is almost entirely interested in the actual prediction and whether it was accurate or not. This analysis, however, also tries deriving implications to contribute to the field of survey methodology. If not only the target gets predicted using complicated algorithms, but already the missing input data gets predicted itself, this double-prediction might make interpretation even more complicated and make it harder to draw insights for survey methodologists.

The idea of this chapter can be summarised as a competition of five different algorithms which each utilise four different models in each of the three countries to predict the binary outcome and chose the single best performing model-algorithm combination. In other words, this chapter compares the performance of 20 different model-algorithm combinations per country to investigate the feasibility of a first contact success prediction.

To explore this idea, the research questions, chapter-specific methods and input models will be introduced in more detail before the results of the analyses are presented in Section 5.3. Lastly, their predictive performance will be compared and discussed to find the best model-algorithm combination.

5.1. Research Questions

This chapter tries to investigate the following six research questions:

1. Conditional on country: which, if any, performance differences are found between the algorithms?
2. Conditional on algorithm: which, if any, performance differences are found between the countries?
3. Which model-algorithm combination best predicts contact in each specific country?
4. Do machine learning algorithms outperform logistic regression predictions in terms of their ability to predict successful first contact?
5. What are the differences in computational time between the algorithms?

6. Does the predictive performance increase when using larger models with a high number of variables but smaller sample size, or when using smaller models which feature a higher sample size?

5.2. Chapter-Specific Methods

To extend the analyses from Chapter 4, the same inclusion and exclusion criteria were used. This means that both respondents and nonrespondents are included in the forthcoming analysis if data on the interviewer contact information and the outcome of the first visit is available.

The following analysis makes use of the same variables as defined in Chapter 4 and applies the same operationalisation as shown in Section 3.3. The only exception is that the variable measuring whether a telephone number was available or not will not be used for the subsequent analyses, since the sample population for this variable differed between Germany and the UK and because the variable was not available in the French sample at all (see also Section 4.2). The outcome variable is defined identically to the outcome variable in Chapter 4 as operationalised in Section 3.3, with no-contact being the negative (0) and contact being the positive outcome (1).

As described above, four different input models are defined. Model 1 uses the full number of possible correlates of contact derived from the literature and as operationalised in Chapter 3.3, including a high number of variables that are only available for respondents. Model 2 begins to reduce this number of substantial variables and simultaneously utilises more of those variables that have a larger proportion of non-missing data. Model 3 drops all variables from the substantive part of the interview entirely and relies only on the paradata and available interviewer data. Lastly, Model 4 also drops the available interviewer data and only includes variables from the contact information sheet. Thus, Model 4 is the smallest model in terms of variables but has the largest sample size.

One of the aims of this analysis is to find out whether Model 1, which features a lot of possible predictors with a reduced sample size or Model 4, which features fewer predictor variables but derives the predictions from a substantially increased sample, performs better or if any of the models in between can unite the benefits of the trade-off.

The following Table 15 shows which predictors make up the different models. While Model 1 contains all 20 variables, Model 2 only features 17 variables, Model 3 contains 11 variables and Model 4 only makes use of 7 variables.

Model 1	**Model 2**	**Model 3**	**Model 4**
Day of the first visit	Day of the first visit	Day of the first visit	Day of the first visit
Hour of the first contact attempt	Hour of the first contact attempt	Hour of the first contact attempt	Hour of the first contact attempt
Type of house	Type of house	Type of house	Type of house
Physical condition of building	Physical condition of building	Physical condition of building	Physical condition of building
Access impediments	Access impediments	Access impediments	Access impediments
Amount of litter in immediate vicinity	Amount of litter in immediate vicinity	Amount of litter in immediate vicinity	Amount of litter in immediate vicinity
Amount of vandalism in immediate vicinity	Amount of vandalism in immediate vicinity	Amount of vandalism in immediate vicinity	Amount of vandalism in immediate vicinity
Interviewer sex	Interviewer sex	Interviewer sex	-
Interviewer first contact workload	Interviewer first contact workload	Interviewer first contact workload	-
Interviewer completion rate	Interviewer completion rate	Interviewer completion rate	-
Interviewer age	Interviewer age	Interviewer age	-
Respondent’s main activity in last seven days	Respondent’s main activity in last seven days	-	-
Degree of urbanicity	Degree of urbanicity	-	-
Respondent sex	Respondent sex	-	-
Number of household members	Number of household members	-	-
Respondent’s education years	Respondent’s education years	-	-
Respondent’s age	Respondent’s age	-	-
Respondent’s marital status	-	-	-
Household income decile	-	-	-
Whether or not a child lives at respondent’s home	-	-	-

Table 15: Model Description

To pre-process the models, all variables with a higher share of missing values than 80% were excluded from the analysis. Machine learning best-practices further recommend excluding 'Near-Zero-Variance'[14] (NZV) variables. NZV variables only carry a limited amount of information because their distribution resembles a constant. However, this dataset contains variables that are of particular interest from a survey methodological perspective. Unfortunately, these variables can have a very skewed distribution with only rare occurrences of events (e.g., units who live in areas with a high amount of vandalism in the immediate vicinity), and thus a low – or NZV – variance distribution. Further recommendations suggest avoiding multicollinearity and thus excluding variables that have a high correlation with each other in the model. As the correlation matrices in Section 4.2 show, there are only few high correlations between the predictors. As all variables can be of substantial methodological importance, it was decided that NZV-variables and those with a multicollinearity were not excluded from the analysis.

To estimate whether these decisions have a substantive impact on the results, a Proof-of-Concept (PoC) analysis for all models using the UK dataset was conducted. The analysis showed that in the case of the first model nine out of a total of 43 factor levels should have been excluded, because they fulfilled the NZV criteria. As expected, these exclusions would particularly affect variables with small cell sizes that are of particular interest from a methodological perspective (for example being sick or amount of litter in the area). Another

[14] A variable is defined as Near-Zero-Variance when the proportion of unique values in its sample distribution is less than 10%.

three factor levels should have been excluded due to a correlation of above 75%. Overall, the Proof-of-Concept analysis showed that the evaluation matrices of the models where NZV and highly correlated variables were excluded only differed slightly in their performance from those models where they were included. The UK Model 1 PoC analysis excluding NZV and highly correlated variables, showed an increase in AUC of 6 in the random forest model, +2 in the XGB, -2 in SVM and -2 in the GLMNET compared to the AUC performances of the UK Model 1, which included NZV and highly correlated variables. The PoC analysis revealed that excluding NZV and highly correlated variables leads to small increases in the algorithms' performances. However, this increase needs to be leveraged against a sacrifice in substantial survey methodological information by excluding interesting variables. Thus, it was decided to loosen the best-practices criteria to a particularly small extent and to retain NZV and highly correlated variables, especially since the performance gains are marginal.

The dataset for the main analysis was split into a 70% training and a 30% testset and predictor variables were centred and scaled to range from 0 to 1 as pre-processing steps following machine learning best practices. 5x5 adaptive cross-validation with random hyperparameter search was chosen as a resampling and hyperparameter tuning method for all models in all countries (see also Chapter 3.9.4)[15]. The following results are based on the application of random forests, XGB, SVM, GLMNET and logistic regression models as introduced in Chapter 3.9.6.

[15] Note that a 5x5 repeated cross-validation is different from a *k*-fold-cross-validation with *k* = 25 discussed earlier. While the former resamples five different samples five times the latter takes 25 samples only once. The former is thus less prone to errors.

5.3. Results

Table 16 shows the characteristics for each model by each country. The overall sample size for each model per country increases noticeably between Model 1 and Model 4 because the included predictor variables are reduced steadily from 20 predictors in Model 1 to eight predictors in Model 4. The complete cases Model 1 in the UK and France features 400 and 402 observations, respectively, whereas Model 1 in Germany only features 228 cases. Model 4 for the German sample consists of 7,866 observations, while the UK sample consists of 4,797 and the French of 4,207 cases. The absolute sample sizes for each country-model test and trainset as well as the respective proportion of successful first contacts are presented for clarity.

The sample characteristics show that the target variable is approximately equally distributed in the train and test datasets. However, it is important to observe that there is an overall lower share of contacts in Model 4 for all countries. This is likely due to the much larger sample size, which was achieved by only using paradata and thus including all observations, regardless of whether they participated in the survey or not. Including all observations from the gross sample that did not participate decreases the overall share of contacts in the dataset.

	United Kingdom				**Germany**				**France**			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
N	400	1,761	1,786	4,797	228	2,187	2,201	7,866	402	1,892	1,932	4,207
% of contacts in N	53.50	54.91	54.87	45.77	73.68	72.97	73.05	66.17	51.74	52.53	52.58	42.21
Test-n	120	529	536	1,440	69	657	661	2,360	121	568	580	1,263
% of contacts in test-n	52.50	52.74	54.29	45.54	71.01	71.99	74.28	66.14	52.66	52.11	53.79	41.25
Train-n	280	1,232	1,250	3,357	159	1,530	1,540	5,506	281	1,324	1,352	2,944
% of contacts in train-n	53.92	55.84	55.12	46.31	74.84	73.39	72.53	66.18	49.58	52.71	52.07	42.62
Variables in model	20	17	10	8	20	17	10	8	20	17	10	8

Table 16: Model Characteristics by Country

Confusion matrices and evaluation statistics for Model 1 are given in Table 17 and Table 18, respectively. Support Vector Machines yielded the highest accuracies compared to the other algorithms in all countries. In the UK the SVM achieved an accuracy of 55.0%, in the French sample 53.7% and the SVM applied on the German data yielded the highest accuracy of all Model 1 algorithms of 71.0%. No accuracy is significantly different from the no-information rate (NIR). Over all algorithms and countries, the true positive rate is higher than the specificity and in some cases the algorithm predicts all cases to be positive at the cost of a non-existing specificity. The SVM yields the highest AUC values in both the UK and French sample (52.6% and 54.0%, respectively). Both predictions from the GLM and logistic regression yielded an AUC of 50.8% in the German sample. With the exception of the logistic regression, the GLM took the least amount of time to compute between 45 to 67 seconds. XGB algorithms on the other hand were computationally most expensive with computing times ranging from roughly 30 to 40 minutes. Compared to the other machine learning algorithms GLMs provided the best AUC/time ratio for all countries.

The results from Table 18 suggest that the accuracy is not a good evaluation parameter for Model 1 in all countries since it does not exceed the NIR significantly. It is also striking that the AUC rarely exceeds the 50% threshold and even in the best model only predicts 54.0% correctly (France, SVM), while in a lot of cases a random guess of contact success would result in a (far) better prediction. Although the results are only slightly better than a random guess, the computation time already exceeded the half-hour mark in some cases. It seems that making use of a lot of substantial variables, which decreases the sample size to train the algorithms, does not result in satisfactory outcomes.

			Predicted								
			United Kingdom			**Germany**			**France**		
			No-contact	**Contact**	**Total**	**No-contact**	**Contact**	**Total**	**No-contact**	**Contact**	**Total**
Observed	**Forest**	No-contact	19	38	57	0	20	20	19	42	60
		Contact	28	35	63	1	48	49	23	37	61
		Total	47	73	120	1	68	69	42	79	121
	XGB	No-contact	22	35	57	1	19	20	26	35	60
		Contact	32	31	63	9	40	49	27	33	61
		Total	54	66	120	10	59	69	53	68	121
	GLM	No-contact	11	46	57	4	16	20	20	41	60
		Contact	9	54	63	9	40	49	27	33	61
		Total	20	100	120	13	59	69	47	74	121
	Logit	No-contact	27	30	57	4	16	20	20	41	60
		Contact	32	31	63	9	40	49	27	33	61
		Total	59	61	120	13	56	69	47	74	121
	SVM	No-contact	3	54	57	0	20	20	14	47	60
		Contact	0	63	63	0	49	49	9	51	61
		Total	3	117	120	0	69	69	23	98	121

Table 17: Model 1 Confusion Matrices per Country

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	45.00	44.17	54.17	48.33	55.00	69.57	59.42	63.77	63.77	71.01	46.28	48.76	43.80	43.80	53.72
p-value *(accuracy > NIR)*	0.958	0.972	0.392	0.842	0.324	0.660	0.675	0.925	0.925	0.560	0.841	0.675	0.939	0.842	0.262
Sensitivity *(true positive rate)*	55.56	49.21	85.71	49.21	100.00	98.96	81.62	81.63	81.63	100.00	61.67	55.00	55.00	55.00	85.00
Specificity *(true negative rate)*	33.33	38.86	19.30	47.37	5.26	0.00	5.00	20.00	20.00	0.00	61.15	42.62	32.79	32.79	22.95
AUC	44.44	43.90	52.51	48.29	52.63	48.98	43.32	50.82	50.82	50.00	46.41	48.81	43.89	43.89	53.96
Time in seconds	639	1,921	45	< 1	589	293	2,423	56	< 1	352	434	1,877	67	< 1	594

Table 18: Model 1 Evaluation Statistics by Country

Table 19 and Table 20 show confusion matrices and evaluation statistics for the second model. Predictions from a logistic regression achieved the highest accuracy for the UK data (58.8%), while random forests produced the highest accuracy for the German and French sample (72.8% and 61.3%, respectively). While none of the algorithms' predictions led to a significant difference from the NIR in Germany, the accuracy of random forests, GLM and logistic regression was significantly higher than the NIR in the UK and the accuracy of all algorithms was higher than the NIR for the French data. Logistic regression prediction also yielded the highest AUC in the UK sample. An AUC of 57.7% was observed from a XGB in the German data and 61.0% from random forests in the French sample. Compared to the other machine learning algorithms GLMs again provided the best AUC/time ratio for all countries.

While for the German data, the predicted accuracy of all algorithms was not better than the NIR, the accuracy of all algorithms was significantly better than the NIR in France. In the UK sample random forests, GLM and logistic regression accuracies were better than the NIR. This means that in these cases, taking the accuracy as an evaluation metric might be useful, since it classifies the outcome correctly at least in these instances. However, it appears that this is not an unambiguous finding, since the utility of the accuracy as an evaluation metric might change depending on the algorithm that is used (as can be seen in the UK), or the input data used (as observed in the differences between France and Germany). Especially in Germany, the algorithms tended to predominantly classify observations into the majority positive class (contact), which explains the high accuracy of these algorithms. However, the resulting high sensitivity is best contextualised by including the specificity. Thus, interpreting the AUC is also useful for all these models. In comparison to Model 1, all AUC values are at least equal to the 50% threshold. In some cases, the results even show a largely improved prediction compared to a random guess. In both the UK and German data, predictions that are about 7 percentage-points better than a random guess can be made using the Model 2 variable set and logistic regression or XGB prediction, respectively, while even a 11 percentage-point better prediction can be observed using random forests in the French data. Despite the medium sample size this improvement already comes at the cost of increasing computational time that varied between 24 minutes for the XGB in Germany and 30 minutes in France. Not only did the UK prediction from a logistic regression yield the highest AUC value of all algorithms in all models, but this also came at the lowest computational cost with the almost instant logistic regression prediction.

Table 21 and Table 22 present the confusion matrices and evaluation statistics for the third model. SVM yielded the highest prediction accuracy for all countries. However, in the UK and German data no algorithm’s accuracy was significantly different from the NIR. In France the accuracies of the random forest, XGB and SVM were significantly higher than the NIR. In the UK and French sample SVMs produced the highest AUC value (52.1% and 58.2%, respectively), while for the German data XGB turned out to be the most successful algorithm with regards to the AUC value. Both SVMs for the UK and France needed about 49 minutes to be computed, while the XGB for Germany needed about 27 minutes.

As with Model 2, it does not appear to be useful to interpret the accuracy for Model 3, since its importance is too susceptible to selection effects. The AUC values for the UK data are only slightly different from a random guess and achieve 52% at most using a SVM approach, which is already computationally demanding despite the relatively small sample size. Model 3 thus does not appear to be useful for predictions in the UK data. The algorithms produce better AUC values in the French data. Even in the worst case the AUC value of a GLM reaches 56% and the computationally expensive SVM even accomplishes an AUC value of 58%. It appears that Model 3 can be of better use in the French data compared to the UK results. Results from the German dataset are mixed. While the XGB reaches the highest AUC value of all models so far (61%), GLM, logistic regression and SVM predictions produce only marginally better results than a random guess and are thus not suitable for any practical application. The average computational costs for the XGB also promote the use of this algorithm. When comparing the gained AUC per second of computational time the random forest proved to be slightly advantageous but also has a lower AUC value overall.

			Predicted								
			United Kingdom			**Germany**			**France**		
			No-contact	**Contact**	**Total**	**No-contact**	**Contact**	**Total**	**No-contact**	**Contact**	**Total**
Observed	**Forest**	No-contact	94	156	250	26	158	184	146	126	272
		Contact	79	205	279	21	452	473	94	202	296
		Total	168	361	529	47	610	657	240	328	568
	XGB	No-contact	93	157	250	47	137	184	154	118	272
		Contact	85	194	279	48	425	473	125	171	296
		Total	178	351	529	95	562	657	279	289	568
	GLM	No-contact	81	169	250	0	184	184	124	148	272
		Contact	55	224	279	0	473	473	100	196	296
		Total	136	393	529	0	657	657	224	344	568
	Logit	No-contact	90	160	250	4	180	184	129	143	272
		Contact	58	221	279	2	471	473	103	193	296
		Total	148	381	529	6	651	657	232	336	568
	SVM	No-contact	16	234	250	0	184	184	122	104	226
		Contact	5	274	279	0	473	473	150	192	342
		Total	21	508	529	0	657	657	272	296	568

Table 19: Model 2 Confusion Matrices per Country

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	56.52	54.25	57.66	58.79	54.82	72.75	71.84	71.99	72.30	71.99	61.27	57.22	56.34	56.69	53.72
p-value *(accuracy > NIR)*	0.044	0.257	0.013	0.002	0.180	0.349	0.554	0.519	0.450	0.519	0.000	0.008	0.024	0.015	0.070
Sensitivity *(true positive rate)*	73.48	69.53	80.29	79.20	98.21	95.56	89.85	100.00	99.57	100.00	68.24	57.77	66.22	65.52	85.00
Specificity *(true negative rate)*	37.60	37.20	32.40	36.00	6.40	14.13	25.54	0.00	2.17	0.00	53.68	56.62	45.59	47.43	22.95
AUC	55.54	53.37	56.34	57.61	52.30	54.85	57.70	50.00	50.88	50.00	60.96	57.19	55.90	56.31	54.86
Time in seconds	1,737	2,103	91	< 1	3,287	2,142	1,463	83.23	1	2,994	1,775	3,010	108	< 1	2,939

Table 20: Model 2 Evaluation Statistics by Country

			Predicted								
			United Kingdom			**Germany**			**France**		
			No-contact	**Contact**	**Total**	**No-contact**	**Contact**	**Total**	**No-contact**	**Contact**	**Total**
Observed	**Forest**	No-contact	60	185	245	48	122	170	149	119	268
		Contact	63	228	291	47	444	491	126	186	312
		Total	123	413	536	95	566	661	275	305	580
	XGB	No-contact	118	127	245	61	109	170	156	112	268
		Contact	132	159	291	68	423	491	133	179	312
		Total	250	286	536	129	532	661	289	291	580
	GLM	No-contact	24	221	245	0	170	170	132	136	268
		Contact	24	267	291	0	491	491	118	194	312
		Total	48	488	536	0	661	661	250	330	580
	Logit	No-contact	65	180	245	1	169	170	133	135	268
		Contact	66	225	291	1	490	491	118	194	312
		Total	131	405	536	2	659	661	251	329	580
	SVM	No-contact	45	200	245	4	166	170	136	132	268
		Contact	41	250	291	2	489	491	107	205	312
		Total	86	450	536	6	655	661	243	337	580

Table 21: Model 3 Confusion Matrices per Country

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	53.73	51.68	54.29	54.10	55.04	74.43	73.22	74.28	74.28	74.58	57.76	57.76	56.21	56.38	58.79
p-value *(accuracy > NIR)*	0.619	0.895	0.517	0.552	0.381	0.485	0.749	0.520	0.520	0.449	0.030	0.030	0.130	0.113	0.008
Sensitivity *(true positive rate)*	78.35	54.64	91.75	77.32	85.91	90.43	86.15	100.00	99.79	99.59	59.62	57.37	62.18	62.18	65.71
Specificity *(true negative rate)*	24.49	48.16	9.79	26.53	18.37	28.24	35.88	0.00	0.50	2.35	55.60	58.21	49.25	49.63	50.75
AUC	51.43	51.40	50.77	51.93	52.14	59.33	61.02	50.00	50.19	50.97	57.61	57.79	55.72	55.90	58.23
Time in seconds	1,188	2,044	71	< 1	2,928	1,344	1,640	93	< 1	3,018	1,284	2,225	62	< 1	2,947

Table 22: Model 3 Evaluation Statistics by Country

Confusion matrices and evaluation statistics for Model 4 are presented in Table 23 and Table 24. In the UK and in France the random forest models predicted all outcomes to be no-contacts, while in Germany they predicted all observations to be contacts. In the UK and French sample, the other algorithms predicted more diversely, but for the German data all algorithms predominantly only predicted all cases to be contacts. Consequently, no algorithm's accuracy for the German data was significantly higher than the NIR. The same outcome can be observed for the French data. Even though the algorithms produced more varied results, the algorithm's accuracy is slightly higher but also not significantly better than the NIR. In the UK sample the logistic regression predictions achieves a significantly higher accuracy than the NIR (*p*-value = 0.0183). While the AUC values for Germany remain close to the 50% threshold, the logistic regression in the UK achieves 54.9% and the XGB in France 57.0%. The best AUC value for the UK can be observed almost immediately (0.04 seconds), while the XGB model in France needed 43 minutes to compute.

Model 4's feasibility is vastly different for the three countries. Using the German data, the algorithms almost unanimously predicted the majority class only, which leads to an inflated sensitivity while underrating the specificity. While marginally better AUC values of 55% and 57% were found for the UK and France, the Model 4 input variables did not lead to viable predictions in the German sample, since the AUC values remained close to the ones of a random guess. Although increases in the AUC values for the UK and French data were found, it seems that the increases in sample size did not lead to unambiguous increases in the AUC overall.

			Predicted								
			United Kingdom			**Germany**			**France**		
			No-contact	**Contact**	**Total**	**No-contact**	**Contact**	**Total**	**No-contact**	**Contact**	**Total**
Observed	**Forest**	No-contact	757	16	773	0	799	799	742	0	742
		Contact	650	17	667	0	1,561	1,561	521	0	521
		Total	1,407	33	1,440	0	2,360	2,360	1,263	0	1,263
	XGB	No-contact	525	248	773	22	777	799	534	208	742
		Contact	406	261	667	33	1,528	1,561	302	219	521
		Total	931	509	1,440	55	2,305	2,360	836	427	1,263
	GLM	No-contact	670	103	773	0	799	799	575	167	742
		Contact	548	119	667	0	1,561	1,561	341	180	521
		Total	1,218	222	1,440	0	2,360	2,360	916	347	1,263
	Logit	No-contact	587	186	773	1	798	799	573	169	742
		Contact	441	226	667	0	1,561	1,561	340	181	521
		Total	1,028	412	1,440	1	2,359	2,360	913	350	1,263
	SVM	No-contact	623	516	773	0	799	799	640	102	742
		Contact	150	151	667	0	1,561	1,561	399	122	521
		Total	773	667	1,440	0	2,360	2,360	1,039	224	1,263

Table 23: Model 4 Confusion Matrices per Country

	United Kingdom					Germany					France				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	53.75	54.58	54.79	56.46	53.75	66.14	65.68	66.14	66.19	66.14	58.75	59.62	59.78	59.70	60.33
p-value *(accuracy > NIR)*	0.489	0.254	0.020	0.018	0.489	0.509	0.692	0.509	0.492	0.509	0.512	0.274	0.237	0.255	0.132
Sensitivity *(true positive rate)*	2.54	39.13	17.84	33.88	22.64	100.00	97.88	100.00	100.00	100.00	0.00	42.03	34.55	34.74	23.42
Specificity *(true negative rate)*	97.93	67.92	86.67	75.94	80.60	0.00	2.70	0.00	0.12	0.00	100.00	71.97	77.49	77.22	86.25
AUC	50.24	53.52	52.25	54.91	51.62	50.00	50.32	50.00	50.06	50.00	50.00	57.00	56.02	55.98	54.83
Time in seconds	2,166	1,383	89	< 1	8,622	4,027	1,603	118	< 1	26,060	2,089	2,602	104	< 1	6,802

Table 24: Model 4 Evaluation Statistics by Country

Table 25 summarises the best results for the algorithms with the highest AUC-value per model and country. XGB and SVMs led to the highest AUC-values in four of twelve cases each, logistic regression in three cases and random forests in one case. The best result for Model 1 was obtained in France with an AUC of 54.0. Model 2 showed higher AUC results than Model 1 in all three countries. For the German sample, the highest of all results was found in Model 3. On the contrary, Model 4 led to the lowest AUC value in the German sample but yielded valuable AUC-values for the French and the second highest values for the UK data. The results presented in Table 25 are depicted in Figure 14 to Figure 17 below, with one figure for each model, displaying the ROC curves for the best algorithm in each country. Figure 14 reveals how close most Model 1 algorithms are to the 50% threshold. Figures 15 and 16 on the other hand show the improvements in AUC, which decrease in Figure 17.

Model	United Kingdom		Germany		France	
	AUC %	Algorithm	AUC %	Algorithm	AUC %	Algorithm
1	52.63	SVM	50.82	GLM/logit	53.96	SVM
2	57.61	logit	57.70	XGB	60.96	Forest
3	52.14	SVM	61.02	XGB	58.23	SVM
4	54.91	logit	50.32	XGB	57.00	XGB

Table 25: Algorithms with the highest AUC Value per Model by Country

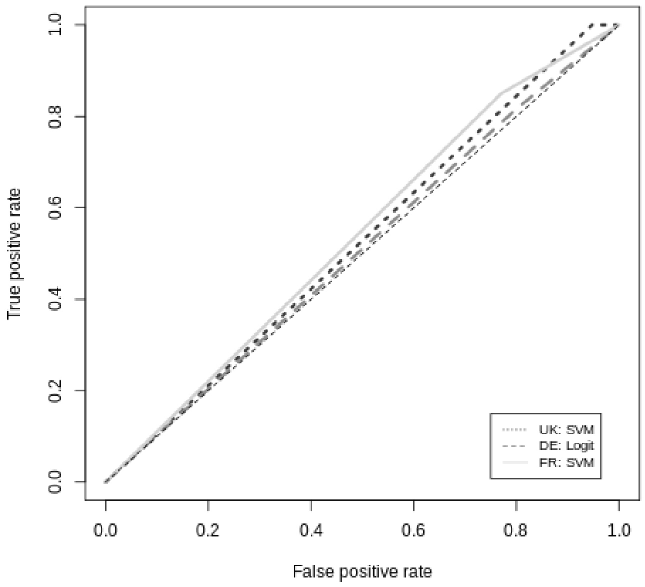

Figure 14: Model 1 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms.

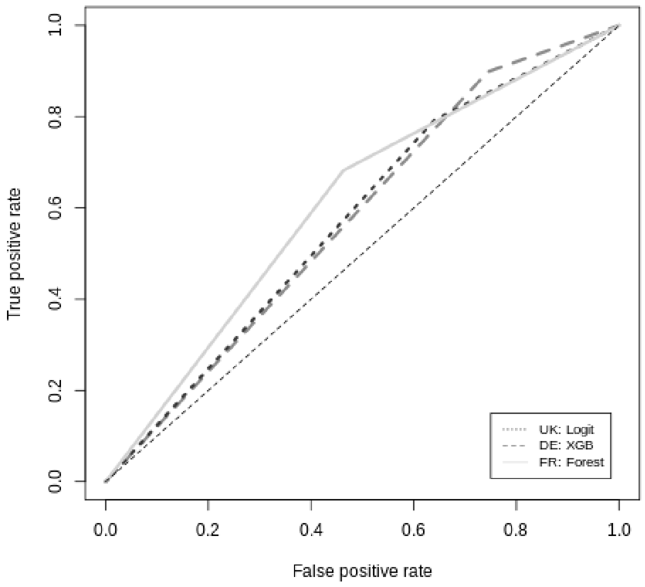

Figure 15: Model 2 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms.

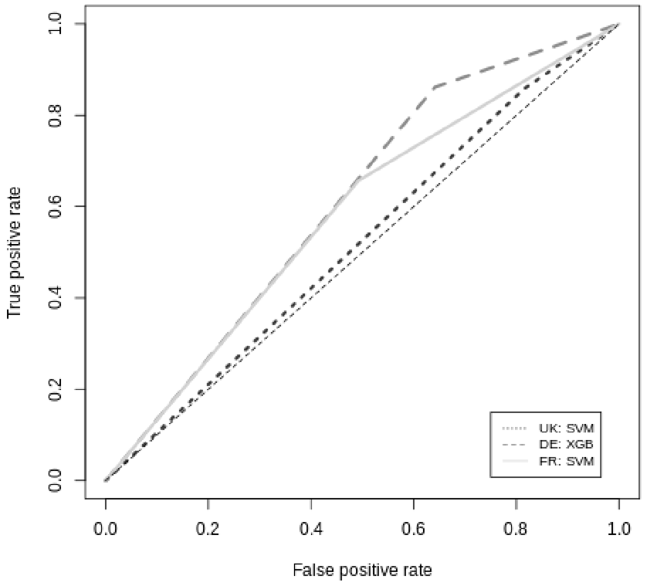

Figure 16: Model 3 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms.

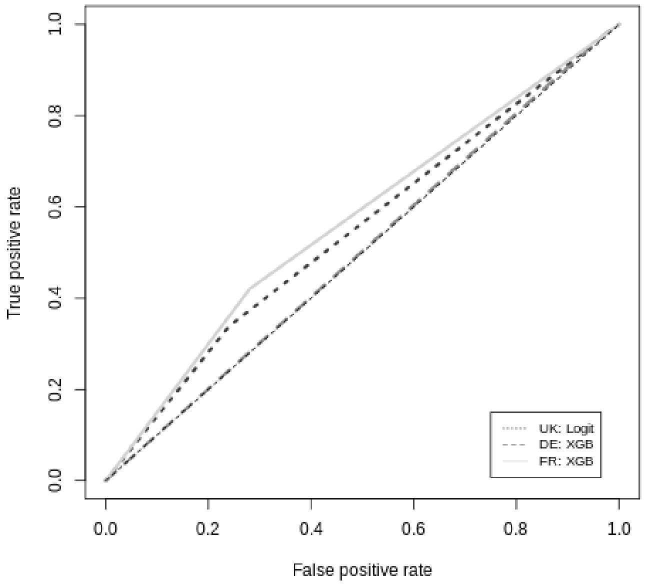

Figure 17: Model 4 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms.

5.4. Discussion & Conclusion

The results show that there is not 'one algorithm to rule them all', but instead that different algorithms need to be compared to find the most efficient and effective prediction for each dataset, variable input model and country. While Model 1 was already about 4 percentage-points better than a random guess for the French data, its predictions were less successful in the UK and impractical in the German data, because the predictions were only 0.8 percentage-points better than a random guess. This marginal predictive improvement does not justify the effort that is needed to produce the prediction. Model 2 however, showed promising results in all three countries with above 7 percentage-points better prediction than guessing in the UK and Germany and almost 11 percentage-points in the French data. A similarly high result was found in Model 3 for the German data with roughly 11 percentage-points better predictions than guessing. Predictions for France in Model 3 were also useful, however Model 3 was almost impractical for the UK data. On the other hand, Model 4 was essentially not better than a random guess for Germany but yielded the second highest AUC-values for the UK and French data.

Consequentially, the answer to both Research Questions 1 (*'Conditional on country: which, if any, performance differences are found between the algorithms?'*) and 2 *Conditional on algorithm: which, if any, performance differences are found between the countries?'*) is that there are both performance differences between the algorithms and countries. There are large differences in the performances of the different algorithms depending on the variable input set.

Within a specific country, a specific algorithm that performs very well on one specific model, might turn out to be useless when applied to a different input model.

The decision of which approach was the most successful depends on the point of view. First, if one needed to choose the *one* input variable set for *all* countries that leads to the highest AUC values possible, this would be Model 2. Second, if one were to select the variable set that led to the highest AUC value for *each* country, this would be Model 2 for the UK and France but Model 3 for Germany. If the aim is to find the one algorithm that produces the most similar results between the countries over all models, this would lead to applying a random forest on the input variable set of Model 4, which has a standard deviation in AUC of only 0.13 between the countries. However, this model-algorithm combination is completely impractical in all countries, since it does not produce any better predictions than a random guess. Lastly, if the aim was to find the one model-algorithm combination that was the most successful over all countries, this would be XGB for Model 3 in Germany. This shows how differently the results can be interpreted.

Research Question 3 (*'Which model-algorithm combination best predicts contact in each specific country?'*) tries to answer the question of which model-algorithm combination is most suitable to predict contact success in each country: Model 2 logistic regression prediction is chosen for the UK (AUC = 57.6), for Germany Model 3 XGB (AUC = 61.0) prediction and for France Model 2 random forest (AUC = 61.0) prediction is chosen for further inspection.

Between the countries the feasibility of logistic regression predictions varies. While in France logistic regression predictions never yielded the highest AUC values regardless of the model, a logistic regression prediction was the most successful algorithm for the German Model 1 dataset, even though the prediction was practically not useful For the UK data the logistic regression prediction outperformed all other algorithms in Model 2, 3 and 4 and even yielded the best AUC value for Model 2 of all compared model-algorithm combinations. In fact, the AUC value of 58% in the UK Model 2 logistic regression prediction was the fifth highest of all compared twelve combinations. Thus, to answer Research Question 4 (*'Do machine learning algorithms outperform logistic regression predictions in terms of their ability to predict successful first contact?'*) it can be said that while logistic regression predictions did not strictly outperform the other algorithms, these machine learning algorithms also did not strictly outperform logistic regression prediction. Instead, it is important to compare all algorithm-model combinations and consider country differences.

When looking at the computational efficiency to answer Research Question 5 (*'What are the differences in computational time between the algorithms?'*), there are obvious differences and there is a clear winner when it comes to computational time. Since the logistic regression prediction did not need any form of machine learning pre-processing, no cross-validation or complex training processes, the results were computed almost instant regardless of the number of observations in the input dataset. In all models for all countries, logistic regression prediction was always the fastest. Similarly, GLMNET predictions came second in speed and were computed quickly. This is an expected result as the GLMNET was designed in a way that it resembles the logistic regression characteristics and just applies it using a machine learning frame. Thus, the computation is similarly straight forward, but more resources for data pre-processing, cross-validation and training are needed. Patterns become interesting from the third fastest model onwards: In Model 1, with the smallest sample size of all models, the ranking in computation speed of the algorithms for all countries is identical: random forests are third, support vector machines come fourth and XGB is last. This means that although XGB is specifically designed to be as efficient as possible, it takes a lot of time to compute predictions when applied onto a dataset with only a small number of cases. In fact, it is not only slower than support vector machines, but even slower than random forests, although it is meant to be an improved version of exactly this method. However, with increases in sample size, changes in this pattern occur. In Model 3, with a medium high number of observations, the support vector machines are the slowest, while XGB is still slower than random forests. Only in Model 4 with the most observations it seems that XGB can finally show its strengths and starts to outperform not only support vector machines but also random forest predictions in terms of computation speed.

Thus, while logistic regression and GLMNET strictly outperform the other algorithms with regards to computational processing time, the differences to the computational times of other algorithms depend on the sample size they are dealing with. It seems that random forests are computed faster than both support vector machines and XGB when applied on smaller sample size datasets, but when there is a lot of data, then support vector machines become very slow in computation and are outperformed by both random forests and particularly by XGB algorithms. However, although computational processing speed is desirable, it is usually of less importance since the focus often lies on accurate predictions even if they take longer to compute.

The remaining Research Question 6 (*'Does the predictive performance increase when using larger models with a high number of variables but smaller sample size, or when using smaller models which feature a higher sample size?'*) aimed to investigate whether the success of the predictions depend on the number of variables or the number of observations in the dataset, since these criteria are mutually exclusive when applying listwise deletion and if no imputation is desired. It appears that again there is no clear answer to this question. When averaging the AUC values of the best performing algorithms of each model per country – in other words a row-average – (computed from Table 25), then the average Model 1 AUC of all countries was 53% which is the lowest of all model averages. This suggests that a larger sample size might be more important for to gain higher AUC values than having a larger number of variables. However, Model 4, which had the largest number of observations, also only has an average AUC over all countries of 54%. This suggests that the AUC does not necessarily increase linearly with increases in sample size because of reductions in variables. Models 2 and 3 on the other hand had AUC averages of 59% and 57%, respectively. This means that there seems to be a trade-off between the number of variables that are needed to make predictions as well as the number of observations that are needed to base the predictions on.

The research questions in this chapter were designed to provide a framework for assessing the 'success' of different combinations of models and algorithms, with success being defined primarily in terms of predictive performance measured in AUC values, alongside consideration of other relevant characteristics like the computing speed. Therefore, these algorithms were investigated with regards to various elements of their performance. More than anything, the analysis showed that a prediction of contact success is possible and showed a successful prediction of 61% of the cases in the best scenario.

The practical relevance of a first contact success prediction appears to be highly context depended. There is not one definite input dataset for all countries that performs best, nor is there one best algorithm that should be applied for all models and/or all countries. Additionally, the predictive performance varied considerably both within and between the countries, which also included specific instances where the predictions were practically useless and not even equal to those of a random guess.

In Chapter 5 a machine learning approach was applied to predict the first contact success of units in the ESS Round 9. The analysis was led by six research question that focussed on finding the best performing algorithm per country given a set of different input models. It has been shown that predictive analysis for contact success is feasible when taking country context into account. If the results from the best-case scenario of this analysis were applicable to a real-world fieldwork process, 11 percentage-points more prospective contacts could be categorised before the contact attempt. This carries the potential to markedly improve estimations of fieldwork costs for agencies.

However, most of the remaining model-algorithm combinations did not come close to the 11 percentage-point improvement and some of them even performed worse that a random guess. One possible key explanation for this finding might be that the overall sample sizes were too small when only working with data from one ESS round. Since machine learning algorithms are designed to train models better when they can leverage larger sample sizes, the following Chapter 6 applies the same algorithms on a larger dataset.

6. A Pool Alone Does Not Make a Summer. Investigating Performance Increases in Contact Success Prediction by Pooling ESS Data

In the previous chapter data from the ESS Round 9 was used to predict first contact attempt success by leveraging five algorithms using four different input variable sets in three countries. Although the feasibility of the approach was shown and some important insights were deduced from the results, one possible explanation for the underwhelming performance of multiple of the algorithms was a lack of a sufficiently large number of observations to base the predictions on. Most machine learning algorithms require complete cases for the input variable set or perform listwise observation deletion otherwise. Naturally, the possibility of even a single missing value increases with the number of variables, which are included in an input variable set. Combined with the already relatively low number of observations in surveys (relative to the masses of data for example in process-generated data), this can lead to reduced precision of the prediction. The models from Chapter 5 that used the largest numbers of variables, were only able to utilise 400 complete cases, which corresponds to 7% of the UK net sample. Only 228 cases (2% of the net sample) were available for training in Germany and 402 cases (9% of the net sample) in France. Even the smallest models could only employ a total of 4,797 cases (87% of the net sample) in the UK, 7,866 cases (92% of the net sample) in Germany and 4,207 cases (98% of the net sample) in France. Moreover, it must be remembered that the sample sizes fall further when the data is split into test and trainsets (see Section 3.9.3). This necessary split means that the algorithm is only trained on 70% of the remaining observations, reducing the number of available training observations to 3,358 for the UK, 5,506 for Germany and 2,945 for France even for the smallest model.

Since machine learning algorithms are designed to perform best when using large amounts of data, it will be an informative step to repeat the analyses on expanded datasets, which increase these sample sizes as much as possible. To achieve this goal, the same algorithms and models as in Chapter 5 will be used, but they will be applied to an extended data source. Instead of just analysing data from the ESS Round 9, data from ESS Round 1 to 9 will be pooled to increase the sample sizes and thus aim to increase the predictive power of the algorithms.

The analyses follow the same criteria as for the analyses in Chapter 5. This means that the same inclusion criteria were used, the same operationalisation applies, the models were built in an identical way and identical pre-processing methods were applied. Data from all ESS rounds was downloaded from the ESS website, pre-processed, harmonised, and eventually combined into a single dataset as outlined in Section 3.1.

By combining the data from all nine rounds of the ESS and analysing them without any reference to the round in which they were collected, the observations will be artificially treated as if they were collected at one single point in time. However, 17 years lie between the measurements made in the first and those made in the ninth round. During these nine rounds associations between variables might have changed for various reasons including, but not limited to, major events like the economic crisis in 2008/2009 or other global and regional events as well as changes in attitudes, and the structure of those societies. It is therefore important to take a closer look at how the associations might have changed over the years and to investigate whether the relationships between the variables of interest to first contact success were stable over time.

In a first step the data on the twenty variables, which were already under examination in previous chapters, is analysed for each of the nine rounds in addition to a pooled dataset for each of the three countries. This means that the resulting tables can quickly become overwhelming. Additionally, associations can vary in two ways over time. On the one hand, they might vary in whether they are significant or not in a specific round. On the other hand, any group differences might also vary in the direction of the difference. To avoid overcomplexity, the tables that summarise the bivariate analyses, try to condense the information by focussing on the *p*-value of an association or group mean difference only and using colour coding to indicate the direction of a difference when applicable. For contextualisation it is recommended to inspect the complete tables, which are provided in the respective folder in the referenced GitLab repository (see Section 3.7). In accordance with the methods, which were described in Section 3.6, Chi^2-tests and independent samples *t*-tests are conducted and presented in the results section before they are interpreted and discussed.

6.1. Research Questions

The analyses will begin with an investigation of the relationships between the predictor variables and the target over time before the predictive performance of the algorithms is examined and compared further. The analyses are led by the following research questions:

1. Are the associations between the predictor variables and the outcome stable over time?

2. Does the increased sample size lead to better predictions?

3. How are gains in predictive power related to computational processing time?

6.2. Results

Table 26 shows the samples sizes for the three countries of interest in all ESS rounds. According to correspondence with the Norwegian Centre for Research Data (NSD), the designated archive for the ESS data, data for Germany Round 5 was excluded from this analysis because the crucial paradata for Germany in this particular round was suffering from severe quality concerns. Furthermore, paradata for France Round 1 is also missing, which is why the entire set of data for France for this particular round was excluded from this analysis. For both the UK and Germany a large increase in sample size can be observed when the first and last rounds' sample sizes are compared. While in the UK the sample of ESS Round 1 featured 4,013 observations, ESS 9 in the UK featured 5,850 cases. Similarly, an increase of 50% in gross sample size can be observed for Germany, while the sample sizes for France remained roughly the same over the years. The pooling increased the available net sample size to 39,958 for the UK, 59,356 for Germany and 32,996 for France, which corresponds to an increase of 628% in the UK, 600% in Germany and 674% in France compared to the sample sizes that were available from ESS Round 9 in Chapter 5.

	ESS 1	ESS 2	ESS 3	ESS 4	ESS 5	ESS 6	ESS 7	ESS 8	ESS 9	Pool
United Kingdom										
gross *n*	4,013	4,032	4,752	4,640	4,640	4,520	5,600	5,000	5,850	43,047
excluded	234	577	372	348	203	174	443	378	360	3,089
net *n*	3,779	3,455	4,380	4,292	4,437	4,346	5,157	4,622	5,490	39,958
Germany										
gross *n*	5,796	5,868	5,712	6,716	-	8,904	9,850	9,456	8,695	60,997
excluded	217	234	288	387	-	40	56	95	224	1641
net *n*	5,579	5,634	5,424	6,329	-	8,864	9,794	9,261	8,471	59,356
France										
gross *n*	-	4,400	4,680	4,500	4,000	4,200	4,173	4,300	4,400	34,653
excluded	-	225	244	181	236	258	147	227	139	1,657
net *n*	-	4,175	4,436	4,319	3,764	3,942	4,026	4,073	4,261	32,996

Table 26: Sample Sizes by Country and ESS Round

6.2.1. Relationships Over Time

Table 27 to Table 29 show the *p*-values for Chi²-tests of independence between the respective predictor variable and the dependent variable of whether contact was established (1) or not (0). Values printed in bold indicate that the relationship is significant at the 5%-significance level and that the variables might be associated.

For the UK data only the main activity in the past seven days as well as whether access impediments were present seem to be related to the contact success of a unit in all nine ESS rounds. For Germany, an association between the degree of urbanicity as well as type of housing unit and the target can be observed over all rounds. For the French data, Chi²-tests for the main activity in the past seven days, type of the housing unit, as well as presences of access impediments, were significant over all rounds suggesting an association that is stable across the years. All other categorical predictors in the countries varied over the years in whether a Chi²-test hinted at independence with the dependent variable or not. In some cases, Chi²-tests hinted at the association of a variable with the target in all but one round (for example day of the first visit in Germany), while for others, the test suggested the independence of a variable and the target in all but one round (for example for litter in the immediate vicinity in Great Britain).

	ESS 1	ESS 2	ESS 3	ESS 4	ESS 5	ESS 6	ESS 7	ESS 8	ESS 9	Pool
Weekday of first visit	**0.0244**	**0.0080**	0.6466	0.5767	**0.0007**	0.0528	0.1059	0.1705	**0.0012**	**0.0073**
Domicile, respondent's description	0.6100	0.0527	0.1249	0.8552	**0.0013**	**0.0031**	0.2783	0.0502	**0.0109**	**0.0001**
Type of House	0.2393	**0.0031**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Physical Condition of building	-	-	-	-	0.0860	**0.0079**	0.0609	**0.0001**	**0.0074**	**0.0001**
Access Impediments	**0.0005**	-	-	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Litter in area	0.7844	0.6800	0.0098	0.1845	0.6501	0.1237	0.8721	0.3730	0.3536	0.1861
Vandalism in area	0.2584	0.8670	0.1552	0.6279	0.2919	0.3943	0.0372	0.1370	0.5381	0.7219
Interviewer's Sex	-	-	-	0.3450	0.4781	0.3750	0.1091	0.6937	0.8360	0.7835
Respondent's Sex	0.9354	0.6762	0.1716	**0.0044**	0.9876	0.1473	0.6797	0.2083	0.4210	0.0514
Children at Home	0.0632	0.2703	0.7526	0.0874	0.9706	**0.0111**	0.2470	0.3648	0.7460	**0.0008**
Marital Status	**0.0004**	**0.0002**	**0.0033**	**0.0001**	**0.0013**	**0.0001**	0.0944	0.0343	0.2110	**0.0001**
Main activity last 7 days	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0005**	**0.0001**

Table 27: p-values of Chi²-tests Between Independent and Dependent Variable for the UK

	ESS 1	**ESS 2**	**ESS 3**	**ESS 4**	**ESS 5**	**ESS 6**	**ESS 7**	**ESS 8**	**ESS 9**	**Pool**
Weekday of first visit	**0.0288**	**0.0298**	0.1192	**0.0003**	-	**0.0009**	**0.0001**	**0.0347**	**0.0001**	**0.0001**
Domicile, respondent's description	**0.0001**	**0.0002**	**0.0188**	**0.0001**	-	**0.0068**	**0.0011**	**0.0001**	**0.0015**	**0.0001**
Type of House	**0.0001**	**0.0001**	**0.0001**	**0.0001**	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Physical Condition of building	-	-	-	-	-	**0.0004**	0.2912	0.4610	0.2749	0.0165
Access Impediments	0.6408	-	-	-	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Litter in area	0.3067	0.1312	0.9458	0.6084	-	0.2488	0.2613	**0.0265**	0.5435	**0.0076**
Vandalism in area	**0.0251**	**0.0115**	0.0918	0.3097	-	0.3199	**0.0043**	**0.0014**	0.3880	**0.0001**
Interviewer's Sex	-	-	-	0.9423	-	0.9105	0.2105	0.2909	0.8477	0.9187
Respondent's Sex	0.1028	0.2144	0.4184	0.3631	-	0.3796	0.0978	0.5283	0.5473	0.3653
Children at Home	0.3591	0.4912	0.1021	0.0738	-	0.2637	0.7681	0.7478	0.7527	**0.0118**
Marital Status	**0.0001**	**0.0268**	0.1496	**0.0066**	-	**0.0471**	0.5621	**0.0396**	0.5279	**0.0001**
Main activity last 7 days	**0.0001**	0.3166	**0.0384**	**0.0107**	-	**0.0001**	0.2704	**0.0002**	0.3893	**0.0001**

Table 28: p-values of Chi^2-tests Between Independent and Dependent Variable for Germany

	ESS 1	**ESS 2**	**ESS 3**	**ESS 4**	**ESS 5**	**ESS 6**	**ESS 7**	**ESS 8**	**ESS 9**	**Pool**
Weekday of first visit	-	0.5393	**0.0003**	0.1351	0.6110	**0.0305**	**0.0032**	0.9822	0.7188	0.3268
Domicile, respondent's description	-	0.3750	**0.0001**	**0.0001**	0.2993	**0.0001**	**0.0013**	**0.0005**	**0.0001**	**0.0001**
Type of House	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Physical Condition of building	-	-	-	-	0.9367	0.0615	0.3379	**0.0293**	**0.0001**	**0.0001**
Access Impediments	-	-	-	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Litter in area	-	0.1818	0.6956	0.7367	0.3617	0.1294	0.0824	0.9788	0.3907	**0.0306**
Vandalism in area	-	0.9981	**0.0203**	0.8366	0.4030	0.0589	**0.0011**	0.1525	0.4234	**0.0099**
Interviewer's Sex	-	-	-	0.8782	0.5307	**0.0387**	0.2641	0.6818	0.1316	0.7264
Respondent's Sex	-	0.3564	0.6919	0.6256	0.1661	**0.0235**	0.7668	0.2897	0.4418	0.5103
Children at Home	-	0.8993	0.3958	0.3579	0.5601	0.0709	**0.0012**	0.9661	0.3119	0.8855
Marital Status	-	-	**0.0004**	**0.0001**	**0.0038**	**0.0013**	0.1077	**0.0020**	0.1509	**0.0001**
Main activity last 7 days	-	**0.0001**	**0.0002**	**0.0001**	**0.0001**	**0.0001**	**0.0043**	**0.0001**	**0.0285**	**0.0001**

Table 29: p-values of Chi^2-tests Between Independent and Dependent Variable for France

	ESS 1	ESS 2	ESS 3	ESS 4	ESS 5	ESS 6	ESS 7	ESS 8	ESS 9	Pool
First Contact Hour	**0.0001**	**0.0001**	**0.0014**	**0.0001**	**0.0002**	**0.0172**	**0.0005**	**0.0009**	**0.0001**	**0.0001**
First Contact Workload	0.4893	0.5908	0.0778	0.1134	0.3708	0.7112	0.8923	0.6145	0.1308	**0.0488**
Completion Rate	-	-	-	**0.0006**	**0.0001**	**0.0001**	**0.0001**	**0.0002**	**0.0001**	**0.0001**
Interviewer Age	-	-	-	**-**	0.5079	0.1879	0.1826	**0.0480**	**0.0234**	0.7460
Number of Household Members	**0.0004**	**0.0004**	0.2337	**0.0006**	**0.0095**	**0.0001**	**0.0001**	**0.0036**	**0.0001**	**0.0001**
Years of Education	**0.0014**	**0.0201**	**0.0009**	**0.0008**	**0.0001**	**0.0067**	0.0568	**0.0004**	**0.0281**	**0.0001**
Household Income	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0015**	0.1437	0.4943	**0.0281**	0.7536	**0.0001**
Respondent's Age	**0.0001**	**0.0003**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0003**	**0.0001**	**0.0043**	**0.0001**

Table 30: p-values for t-tests of Mean Differences Between Units Who Were Contacted (1) and Those Who Were Not (0) in the UK

	ESS 1	ESS 2	ESS 3	ESS 4	ESS 5	ESS 6	ESS 7	ESS 8	ESS 9	Pool
First Contact Hour	**0.0000**	0.2268	**0.0002**	**0.0001**	**-**	**0.0001**	**0.0001**	**0.0001**	**0.0600**	**0.0001**
First Contact Workload	0.7097	**0.0139**	**0.0016**	0.0667	-	**0.0016**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Completion Rate	-	-	-	**0.0001**	**-**	**0.0001**	**0.0001**	**0.0001**	**0.0038**	**0.0001**
Interviewer Age	-	-	-	-	-	**0.0002**	**0.0002**	0.8446	0.1585	**0.0001**
Number of Household Members	**0.0001**	0.0679	**0.0001**	**0.0001**	**-**	**0.0078**	**0.0006**	**0.0048**	**0.0001**	**0.0001**
Years of Education	**0.0009**	**0.0001**	**0.0001**	**0.0030**	**-**	**0.0001**	**0.0001**	**0.0209**	**0.0170**	**0.0001**
Household Income	0.0985	0.1463	0.9501	0.9285	-	**0.0271**	0.7713	0.8632	0.3361	0.7449
Respondent's Age	**0.0297**	**0.0067**	0.3863	**0.0033**	-	**0.0001**	**0.0011**	**0.0010**	0.6590	**0.0001**

Table 31: p-values for t-tests of Mean Differences Between Units Who Were Contacted (1) and Those Who Were Not (0) in Germany

	ESS 1	**ESS 2**	**ESS 3**	**ESS 4**	**ESS 5**	**ESS 6**	**ESS 7**	**ESS 8**	**ESS 9**	**Pool**
First Contact Hour	-	**0.0060**	**0.0013**	**0.0083**	0.7826	0.2309	**0.0007**	**0.0004**	**0.0001**	**0.0081**
First Contact Workload	-	0.8118	0.1736	**0.0002**	**0.0304**	0.7343	**0.0002**	**0.0001**	0.8636	0.5420
Completion Rate	-	-	-	**0.0001**	**0.0001**	**0.0002**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
Interviewer Age	-	-	-	-	-	0.1962	**0.0022**	0.5491	0.1548	**0.0103**
Number of Household Members	-	0.8024	**0.0344**	**0.0047**	**0.0014**	**0.0421**	**0.0001**	**0.0058**	**0.0007**	**0.0001**
Years of Education	-	**0.0001**	**0.0001**	**0.0001**	**0.0002**	**0.0001**	**0.0001**	**0.0018**	**0.0038**	**0.0001**
Household Income	-	**0.0055**	0.0697	**0.0050**	0.1691	0.1520	0.4928	0.8596	0.5222	**0.0001**
Respondent's Age	-	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	0.9559	**0.0001**	**0.0185**	**0.0001**

Table 32: p-values for t-tests of Mean Differences Between Units Who Were Contacted (1) and Those Who Were Not (0) in France

Table 30 to Table 32 show the *p*-values for *t*-tests of group mean differences between those units who were successfully contacted at the first attempt and those which were not. Again, *p*-values printed in bold indicate a significant group mean difference. In addition to the coding in bold, the cell background indicates whether contacted units had a higher group mean of the respective variable than uncontacted units or not. If contacted units had a higher group mean average, the mean difference between the groups is negative, indicated by a dark grey colour code. If the group mean of successfully contacted units was lower than those of unsuccessfully contacted units, the group mean difference was positive and is indicated by a light grey colour-coded cell background. Interpreting whether it is substantially important that a difference is positive or negative and what this means practically depends on the specific variable under observation and is also not of particular interest in this analysis. Again, interested readers can find the full tables in the GitLab repository to make use of the results as required. The focus here lies on whether the direction of a group difference changed over the years or remained stable to identify which of the variables are constant correlates of contact over time.

A detailed explanation for some of the relationships clarifies the idea of this analysis: In the UK there was a significant group mean difference between successfully contacted at first attempt and uncontacted at first attempt units for the average first contact hour, indicated by the *p*-value printed in bold, in all rounds (see Table 30). Additionally, this group mean difference was negative in all rounds (indicated by the dark grey colour code background), meaning that successfully contacted units were contacted significantly later in the day regardless of the ESS round. The variables 'Years spent in education' for the UK shows a similar picture: Here, units who were successfully contacted at first attempt spent fewer years in education on average than units who were uncontacted at first attempt, which results in a positive average group mean difference indicated in light grey. This means that units that were contacted at first attempt on average had received slightly less formal schooling in their lives. The consistent light grey background of this line shows that this mean difference has the same direction over all ESS rounds and is thus stable across time, even if it becomes insignificant in Round 7 (indicated by non-bold typesetting). One single colour code per line indicates the stability of the mean difference's direction over time. It might still vary noticeably in magnitude and significance, but the general direction persists.

Besides stable findings, there are also some varying results: In the UK contacted units were visited by interviewers with a higher mean interviewer workload than uncontacted units in Rounds 1, 5, 7 and 9. However, in the remaining rounds units that were successfully contacted at first attempt were visited by interviewers with a lower mean workload than uncontacted units. Lines with more than one colour code indicate that the direction for group mean differences varied and was not stable over time. However, in the case of interviewer first contact workload in the UK, the results remain inconclusive since none of the group mean differences reached statistical significance.

In other cases, this instability in the direction of the group mean differences does matter. Table 30 shows that in the UK ESS Round 8 successfully contacted units were visited by significantly older interviewers, on average, than uncontacted units. This group mean differences flips in UK ESS Round 9 when successfully contacted units are visited by significantly younger interviewers. Cases in which an instability occurs for significant group mean differences are of very great interest for a survey methodologist and are worth investigating in more detail with other, more advanced methods. Unfortunately, this must be up to future research. This analysis can only point towards these instabilities, while explaining them is out of the scope of this thesis. Having introduced the presentation of the tables in detail with these examples, the report in the next paragraphs of Table 30 to Table 32 summarises their highlights more briefly.

In the UK stable and predominantly significant group mean differences between successfully contacted and uncontacted units at first attempt for all rounds can be found in relation to the contact time of the day, interviewer completion rate, number of household members, years of education, household income and respondent’s age. In all rounds contacted units were visited at later hours of the day, by interviewers with a higher completion rate, had more household members, spent slightly fewer years in education on average and were older on average than uncontacted units at first contact attempt. While the stability of the group differences for the age of interviewer remains unclear, group mean differences for interviewer first contact workload remain inconclusive since they do not reach statistical significance.

In Germany successfully contacted units were contacted at later hours of the day by more successful interviewers, spent less years in education and had a larger number of household members. These findings are stable and predominantly significant over all ESS rounds. Contacted units were also predominantly significantly older on average than uncontacted units in all rounds but Round 9. Since the positive group mean difference for respondent's age in Round 9 did not reach statistical significance, while most others did, this might be a hint that either respondent age is also stable over time nonetheless or that this marks the beginning of a change in the relationship. A similar, yet less supported, result can be found for the group difference in interviewer age. Successfully contacted units were visited by significantly older interviewers in Round 6 and 7, by older interviewers in Round 8 (not reaching statistical significance) and by younger interviewers in Round 9 (also not reaching statistical significance). Whether or not the direction of this result is stable over time remains debatable and more measurement points than the four from Round 6 to 9 would be required. Group mean differences in household income deciles are unstable over time and rarely reach statistical significance. An important result can be found for the interviewer workload: while from Round 1 to 4 successfully contacted units tended to be visited by interviewers with a lower workload (significant in Round 2 and 3), this group mean difference flips from Round 5 onwards, when successfully contacted units were contacted by interviewers with significantly higher average first contact workloads. While this seems to be stable since ESS Round 5, it remains unclear what causes this shift and further investigation would be needed to find out whether this group difference has stabilised or not.

In France successfully contacted units were contacted by more successful interviewers, spent less years in education and had a larger number of household members across all rounds, on average. Similar to the results for Germany, contacted units in France were also significantly older on average than uncontacted units in all rounds but Round 7. Since the positive group mean difference for respondent's age in Round 7 did not reach statistical significance, while all others did, it seems reasonable to conclude that the group difference for respondent age is stable over time. The remaining group mean differences remain inconclusive or appear to be unstable over time. While successfully contacted units were contacted at significantly later hours of the day from Rounds 1 to 3, they were contacted at earlier hours of the day from Round 4 onwards (significantly earlier in Rounds 7, 8, and 9). While successfully contacted units are visited by older interviewers on average, this result remains inconclusive in most rounds since the analysis mostly does not reach statistical significance. Successfully contacted units tend to have a lower household income than uncontacted units in all rounds but Round 9. However, the group mean differences are only significant in Round 1 and 3, thus the stability and importance of this variable remains unclear. The results for the interviewer workload vary a lot. While successfully contacted units are visited by interviewers with a typically lower workload in Round 2, 3, 4, 7 and 9 (significant in Round 2 and 7), they were visited by interviewers with a higher workload in Round 5, 6 and 8 (significant in Round 5 and 8). Thus, the stability and importance of the interviewer workload remains questionable.

6.2.2. Prediction of Contact Success

The specification of each model using the pooled data are shown in Table 33. Due to the listwise deletion the number of observations per model increases considerably with a reduction in included variables. The smallest number of observations can be found for Model 1 in Germany (*n* = 2,890), while the largest model contains 31,040 complete cases (Germany, Model 4). In the most extreme case in Chapter 5 the algorithms were only trained on 159 cases (Germany, Model 1). The pooling leads to an increase in available training observations even in the smallest trainset, which can now utilise 2,022 cases (Germany, Model 1). Similarly, the smallest testset in Chapter 5 only contained 69 observations (Germany, Model 1), in contrast to the smallest testset of this chapter using the pooled data, which contains 719 cases (France, Model 1). The absolute sample sizes for each country-model and their respective test and trainsets as well as the respective proportion of successful first contacts can be seen in Table 33.

	United Kingdom				**Germany**				**France**			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
N	3,924	9,423	9,566	22,527	2,890	8,758	8,831	31,040	3,301	7,631	7,698	19,805
% of contacts in N	49.82	52.99	52.98	43.81	63.52	67.20	67.20	60.92	50.07	52.75	52.80	40.73
Test-n	1,107	2,590	2,603	5,228	863	2,611	2,636	9,050	719	1,303	1,340	1,369
% of contacts in test-n	51.58	53.32	53.28	44.03	64.42	66.64	66.16	60.56	50.34	50.57	53.65	41.34
Train-n	2,746	6,596	6,696	15,768	2,022	6,130	6,181	21,728	2,310	5,341	5,388	13,863
% of contacts in train-n	49.27	52.88	52.77	43.70	63.05	67.45	67.65	61.02	49.52	53.47	53.39	40.49
Variables in model	20	17	10	8	20	17	10	8	20	17	10	8

Table 33: Model Characteristics by Country for Pooled ESS Data.

Next, the confusion matrices and evaluation characteristics for all models will be presented in turn. The confusion matrix and model evaluation characteristics for Model 1 are depicted in Table 34 and Table 35. While predictions of contacts and non-contacts were almost evenly distributed in the UK and France across all algorithms, the algorithms predicted non-contact noticeably less often than contact for the German data. Random Forests for the German data even predicted all cases to be contacts and none to be non-contacts (see Table 34). Table 35 underlines this finding. While the accuracy of all models is better than the NIR in both the UK and France, the accuracies are never significantly different from the NIR for the German data. In the UK the best prediction was found for SVMs, which yielded the highest AUC value of 58.6%, closely followed by a logistic regression, which achieved an AUC of 58.4% and a GLM, reaching an AUC value of 58.2%. In Germany, the best prediction was found for the XGB, which yielded an AUC value of 58.1%. In France, the best prediction was found for random forests with an AUC value of 61.1%. Across all countries, random forests took the most time to compute ranging from 63 to 88 minutes, followed by XGB (25 to 29 minutes), SVM (7 to 13 minutes), GLM (2 minutes) and logistic regression, which was computed in less than a second.

In Chapter 5 when only ESS Round 9 data was used, most algorithms in Model 1 did not even exceed the 50% threshold. Although AUC values of some algorithms for the German Model 1 data were not as high as those of France and the UK, the XGB AUC value reached 58%, which is an improvement of 8% points compared to the most successful Model 1 algorithm of Chapter 5. Averaging over all algorithms and countries the AUC values for the

Model 1 input dataset increased to 57% using the pooled data compared to an average AUC of roughly 48% using the ESS Round 9 data only.

			Predicted								
			United Kingdom			Germany			France		
			Non-contact	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**
Observed	**Forest**	Non-contact	277	259	536	0	307	307	219	138	357
		Contact	213	358	571	0	556	556	142	220	362
		Total	490	617	1,107	0	863	863	361	358	719
	XGB	Non-contact	310	226	536	111	196	307	194	163	357
		Contact	249	322	571	111	445	556	144	218	362
		Total	559	548	1,107	222	641	863	338	381	719
	GLM	Non-contact	295	241	536	39	268	307	221	136	357
		Contact	221	350	571	30	526	556	152	210	362
		Total	516	591	1,107	69	794	863	373	346	719
	Logit	Non-contact	290	246	536	71	236	307	217	140	357
		Contact	213	358	571	68	488	556	156	206	362
		Total	536	571	1,107	139	724	863	373	346	719
	SVM	Non-contact	295	241	536	47	260	307	205	152	357
		Contact	216	355	571	51	505	556	149	213	362
		Total	487	571	1,107	98	765	863	354	365	719

Table 34: Model 1 Confusion Matrices per Country for Pooled Data.

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	57.36	57.09	58.27	58.54	58.72	64.43	64.43	65.47	64.77	63.96	61.06	57.30	59.94	58.83	58.14
p-value *(accuracy > NIR)*	0.0001	0.0001	0.0001	0.0001	0.0001	0.5155	0.5155	0.2735	0.4308	0.6267	0.0001	0.0001	0.0001	0.0001	0.0001
Sensitivity *(true positive rate)*	0.6270	0.5639	0.6130	0.6270	0.6217	1.0000	0.8004	0.9460	0.8777	0.9083	0.6077	0.6022	0.5801	0.5691	0.5884
Specificity *(true negative rate)*	0.5168	0.5784	0.5504	0.5410	0.5504	0.0000	0.3616	0.1270	0.2313	0.1531	0.6134	0.5434	0.6190	0.6078	0.5742
AUC	57.20	57.10	58.20	58.40	58.60	50.00	58.10	53.7	55.4	53.10	61.10	57.30	60.00	58.80	58.10
Time in seconds	5,324	1,690	126	< 1	818	3,797	1,740	134	< 1	418	4,197	1,543	123	< 1	523

Table 35: Model 1 Evaluation Statistics by Country for Pooled Data.

The confusion matrix and model evaluation characteristics for Model 2 are shown in Table 36 and Table 37. As with the results from Model 1, Table 36 shows that the algorithms predict both outcome options in the UK and France more evenly than in Germany, where the algorithms predominantly predict the observations to be contacts. In both the UK and France all algorithms perform significantly better than the NIR, while in Germany only the accuracy of the random forest outperforms the NIR significantly. The AUC values of all algorithms are similar across countries. In the UK, GLM reaches the highest AUC of 59.6% but is closely followed by the logistic regression (59.5%) and even the worst algorithm still achieves an AUC value of 58.1% (SVM). XGB yielded the highest AUC value for the German data (57.8%), while SVMs even undercut the 50% threshold and only reached 49.8%. For the French data, SVMs were most performant and reached an AUC value of 58.4%. Computational time varied largely across algorithms but followed an identical ranking between the countries: Random forests took the longest time to compute (167 to 235 minutes). They were followed by SVMs (36 to 60 minutes), XGB (29 to 43 minutes), GLM (3 to 12 minutes) and logistic regression (less than a second).

The reduction in the number of variables from Model 1 to Model 2 led to a large increase in available sample size that the algorithms could be trained on. It becomes apparent that the accuracy of predictions in Germany gets weaker, since only one algorithm reaches an accuracy which is significantly better than the NIR. While most algorithms tend to be able to leverage sensitivity and specificity quite well, there are algorithms for the German data that apparently struggle with predictions: random forests, GLM, logistic regression and SVM almost entirely predict all units to be ‘contacts’, while only XGB distinguishes in more detail. Unsurprisingly, XGB yields the highest AUC value of all Model 2 algorithms for Germany. Although multiple AUC values for France remain above the 57% mark, no algorithm exceeds the highest Model 1 AUC value of 61%. Instead, the highest Model 2 AUC for France is 58%. On the other hand, GLM achieves an AUC of 60% for the UK data, which is the highest value of all compared AUC values for the UK.

Overall, the average AUC value for all models in all countries using the pooled data decreases slightly compared to Model 1 (57% versus 56%) but is still 2 percentage-points higher than the average AUC values of Model 2 in Chapter 5 (54% versus 56%).

			Predicted								
			United Kingdom			**Germany**			**France**		
			Non-contact	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**
Observed	**Forest**	Non-contact	540	669	1,209	105	766	871	295	349	644
		Contact	384	997	1,381	63	1,677	1,740	197	462	659
		Total	924	1,666	2,590	168	2,443	2,611	492	811	1,303
	XGB	Non-contact	540	669	1,209	248	623	871	343	301	644
		Contact	384	997	1,381	225	1,515	1,740	244	415	659
		Total	924	1,666	2,590	473	2,138	2,611	587	716	1,303
	GLM	Non-contact	562	647	1,209	5	866	871	252	392	644
		Contact	378	1,003	1,381	4	1,736	1,740	167	492	659
		Total	940	1,650	2,590	9	2,602	2,611	419	884	1,303
	Logit	Non-contact	571	638	1,209	39	832	871	304	340	644
		Contact	390	991	1,381	40	1,700	1,740	208	451	659
		Total	961	1,629	2,590	79	2,532	2,611	512	791	1,303
	SVM	Non-contact	550	659	1,209	0	871	871	299	345	644
		Contact	405	976	1,381	6	1,734	1,740	195	464	659
		Total	955	1,635	2,590	6	2,605	2,611	494	809	1,303

Table 36: Model 2 Confusion Matrices per Country for Pooled Data.

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	59.34	59.15	60.42	60.31	58.92	68.25	67.52	66.68	66.60	66.41	58.10	58.17	57.10	57.94	58.56
p-value *(accuracy > NIR)*	0.0001	0.0001	0.0001	0.0001	0.0001	0.0421	0.1753	0.4926	0.5257	0.6072	0.0001	0.0001	0.0001	0.0001	0.0001
Sensitivity *(true positive rate)*	0.7219	0.6662	0.7263	0.7176	0.7067	0.9638	0.8707	0.9977	0.9770	0.9966	0.7011	0.6297	0.7466	0.6844	0.7041
Specificity *(true negative rate)*	0.4467	0.5062	0.4648	0.4723	0.4549	0.1206	0.2847	0.0057	0.0448	0.0000	0.4581	0.5326	0.3913	0.4720	0.4643
AUC	58.40	58.60	59.60	59.50	58.10	54.20	57.80	50.20	51.10	49.80	58.00	58.10	56.90	57.80	58.40
Time in seconds	14,113	1,745	739	< 1	3,013	13,310	2,634	703	< 1	3,652	10,061	2,093	188	< 1	2,217

Table 37: Model 2 Evaluation Statistics by Country for Pooled Data.

Table 38 and Table 39 show the confusion matrix and evaluation statistics for Model 3. While the UK and French data yields a fairly even prediction of contact and non-contact, GLM, logistic regression and SVM predict contact almost exclusively for the German dataset. While accuracies for the French data remains significantly better than the NIR for all algorithms but the GLM at a 5% significance level, predictions of accuracy are not significantly better than the NIR in the UK for any algorithms, and only for random forests in the case of Germany. In the UK XGB led to the highest AUC of 53.1%. Both XGB and random forests yielded high AUC values for the German data with 57.6% and 57.7 respectively, while AUC values for GLM, logistic regression and SVM all remained close to the random guess threshold. For the French data even the lowest achieved AUC value still reached 53.6% (SVM), while XBG yielded the highest AUC value of 59.1%. Across all countries the computational time for the algorithms was comparable: Random forests were the slowest, taking 117 to 158 minutes, followed by SVMs (49 to 73 minutes), XGB (28 to 33 minutes), GLM (2 to 11 minutes) and logistic regression, which was again computed in under a second.

With a further reduction in variables, the trends from Model 2 continue in Model 3. Prediction accuracy is not significantly better than the NIR in any of the UK algorithms and even the highest AUC value only reaches 53%, which is just 1 percentage-point higher than the best Model 3 algorithm using just the ESS Round 9 data. AUC values in Germany and France, on the other hand, look promising. They reach values of more than 57% in Germany and even above 59% in France, respectively. Overall, the average AUC value drops to about 54%, because of the poor performance of all algorithms in the UK and the poor performance of GLM, logistic regression and SVMs in Germany, questioning the utility of the pooling approach.

			Predicted								
			United Kingdom			**Germany**			**France**		
			Non-contact	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**
Observed	**Forest**	Non-contact	503	713	1,216	225	667	892	302	319	621
		Contact	506	881	1,387	175	1,569	1,744	243	476	719
		Total	1,009	1,594	2,603	400	2,236	2,636	545	795	1,340
	XGB	Non-contact	581	635	1,216	254	638	892	336	285	621
		Contact	575	812	1,387	227	1,517	1,744	258	461	719
		Total	1,156	1,447	2,603	481	2,155	2,636	594	746	1,340
	GLM	Non-contact	128	1,088	1,216	1	891	892	211	410	621
		Contact	109	1,278	1,387	1	1,743	1,744	192	527	719
		Total	237	2,366	2,603	2	2,634	2,636	403	937	1,340
	Logit	Non-contact	337	879	1,216	3	889	892	242	379	621
		Contact	339	1,048	1,387	4	1,740	1,744	206	513	719
		Total	676	1,927	2,603	7	2,629	2,636	448	892	1,340
	SVM	Non-contact	223	993	1,216	0	892	892	217	404	621
		Contact	207	1,180	1,387	0	1,744	1,744	180	539	719
		Total	430	2,173	2,603	0	2,636	2,636	397	943	1,340

Table 38: Model 3 Confusion Matrices per Country for Pooled Data.

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	53.17	53.52	54.01	53.21	53.90	68.06	67.19	66.16	66.12	66.16	58.06	59.48	55.07	56.34	56.42
p-value *(accuracy > NIR)*	0.5548	0.4146	0.2337	0.5393	0.2714	0.0204	0.1376	0.5091	0.5255	0.5091	0.0007	0.0001	0.1554	0.0258	0.0226
Sensitivity *(true positive rate)*	0.6352	0.5854	0.9214	0.7556	0.8508	0.8997	0.8698	0.9994	0.9977	1.0000	0.6620	0.6412	0.7330	0.7135	0.7497
Specificity *(true negative rate)*	0.4137	0.4778	0.1053	0.2771	0.1834	0.2522	0.2848	0.0011	0.0034	0.0000	0.4863	0.5411	0.3398	0.3897	0.3494
AUC	52.44	53.16	51.33	51.64	51.71	57.60	57.70	50.00	50.10	50.00	57.40	59.10	53.60	55.20	55.00
Time in seconds	9,505	1,998	687	< 1	4,417	8,759	1,501	663	< 1	2,948	7,037	1,739	168	< 1	2,998

Table 39: Model 3 Evaluation Statistics by Country for Pooled Data.

The confusion matrix and evaluation statistics for Model 4 are illustrated in Table 40 and Table 41. Confusion matrices for Model 4 show a different picture than observed for the previous models. In the UK contact gets predicted considerably less often than non-contact. Similarly but more extreme, the algorithms for the French dataset also predict contact less often or even not at all in the cases of random forests and GLM. Interestingly, the within-country pattern for Germany persists but mirrors the results for the UK and French Model 4 results: contact gets predicted much more frequently than non-contact. For the Model 4 variable input dataset the computational time varied largely across the algorithms and the ranking was also slightly different for the countries. While in the UK and France random forests took the most time to be computed (249 and 208 minutes), followed by SVMs (214 and 195 minutes), SVMs were the slowest for the German data and needed 874 minutes to be computed, followed by random forests with 363 minutes. The remaining algorithms followed the same ranking across the countries: XGB came third (35 to 50 minutes), GLM (12 and 14 minutes) and logistic regression (less than a second).

Overall, Model 4 underlines the findings from Model 3: despite a reduction in variables and thus the largest number of available observations, the AUC values are practically useless for the German data, because they never even exceed the 51% mark and even in the best-case scenario only reach 53% for the French data. The average AUC value of all algorithms is 51%, which means that it is 1 percentage-point lower than in the Model 4 equivalent in Chapter 5, which only leveraged the ESS Round 9 data.

			Predicted								
			United Kingdom			**Germany**			**France**		
			Non-contact	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**	**Non-contact**	**Contact**	**Total**
Observed	**Forest**	Non-contact	2,775	151	2,926	156	3,413	3,569	803	0	803
		Contact	2,082	220	2,302	153	5,328	5,481	566	0	566
		Total	4,857	371	5,228	309	8,741	9,050	1,369	0	1,369
	XGB	Non-contact	2,492	434	2,926	122	3,447	3,569	743	60	803
		Contact	1,827	475	2,302	147	5,334	5,481	501	65	566
		Total	4,319	909	5,228	269	8,781	9,050	1,244	125	1,369
	GLM	Non-contact	2,634	292	2,926	112	3,457	3,569	803	0	803
		Contact	1,952	350	2,302	126	5,355	5,481	566	0	566
		Total	4,586	642	5,228	238	8,812	9,050	1,369	0	1,369
	Logit	Non-contact	2,567	359	2,926	206	3,363	3,569	749	54	803
		Contact	1,886	416	2,302	213	5,268	5,481	499	67	566
		Total	4,453	775	5,228	419	8,631	9,050	1,248	121	1,369
	SVM	Non-contact	2,811	115	2,926	126	3,443	3,569	728	75	803
		Contact	2,165	137	2,302	155	5,326	5,481	479	87	566
		Total	4,976	252	5,228	281	8,769	9,050	1,207	162	1,369

Table 40: Model 4 Confusion Matrices per Country for Pooled Data.

	United Kingdom					**Germany**					**France**				
Statistic	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**	**Forest**	**XGB**	**GLM**	**Logit**	**SVM**
Accuracy *(% correctly classified)*	57.29	56.75	57.08	57.06	56.39	60.60	60.29	60.41	60.49	60.24	58.66	59.02	58.66	59.61	59.53
p-value *(accuracy > NIR)*	0.0281	0.1296	0.0545	0.0576	0.2747	0.4789	0.7085	0.6227	0.5644	0.7373	0.5116	0.4030	0.5116	0.2466	0.2643
Sensitivity *(true positive rate)*	0.0956	0.2063	0.1520	0.1807	0.0595	0.9721	0.9732	0.9770	0.9611	0.9717	0.0000	0.1148	0.0000	0.1184	0.1537
Specificity *(true negative rate)*	0.9484	0.8517	0.9002	0.8773	0.9607	0.0437	0.0342	0.0314	0.0577	0.0353	1.0000	0.9253	1.0000	0.9328	0.9066
AUC	52.20	52.90	52.61	52.90	51.01	50.80	50.40	50.40	50.90	50.40	50.00	52.00	50.00	52.60	53.00
Time in seconds	14,946	2,803	867	< 1	12,872	21,793	3,022	837	< 1	52,468	12,533	2,158	751	< 1	11,749

Table 41: Model 4 Evaluation Statistics by Country for Pooled Data.

The best performing algorithms in terms of their AUC value per country are listed in Table 42. XGB produced the highest AUC values in six of the twelve cases, followed by SVM (3), logistic (2), and forests and GLM (both 1). The highest AUC for Model 1 was achieved in France using a random forest and resulting in an AUC of 61.1%, which is also the highest AUC achieved by all algorithms in all countries. The highest AUC value for Model 2 was achieved in the UK using a GLM, while the highest AUC value for Model 3 was achieved in France using an XGB. SVMs in France produced the highest AUC value for all countries using the Model 4 data input and achieved an AUC of 53.0%. The highest AUC over all models for the UK can be found for Model 2 with 59.2%. For both Germany and France Model 1 yielded the highest AUC value of all models with 58.10 and 61.1%, respectively. The results presented in Table 42 are depicted in Figure 18 to Figure 21 below, with one figure for each model, displaying the ROC curves for the best algorithm in each country. Especially Figure 18 and Figure 19 show how close the ROC curves are to one another across the countries. The ROC curves in Figure 20 are already less similar to each other and already it gets visible that the underlying AUC is smaller than the one of the earlier figures. Finally, the ROC curves in Figure 21 show that the AUC areas are the smallest compared to those of the earlier figures and that the ROC curve for Germany is almost identical to the 50% reference line, indicating that this algorithm only performs slightly better than a random guess.

Model	United Kingdom		Germany		France	
	AUC %	Algorithm	AUC %	Algorithm	AUC %	Algorithm
1	58.60	SVM	58.10	XGB	61.10	Forest
2	59.60	GLM	57.80	XGB	58.40	SVM
3	53.16	XGB	57.70	XGB	59.10	XGB
4	52.90	XGB/Logit	50.90	Logit	53.00	SVM

Table 42: Algorithms with the Highest AUC Value per Model by Country for Pooled Data. Highlighted Best Model-Algorithm Combination.

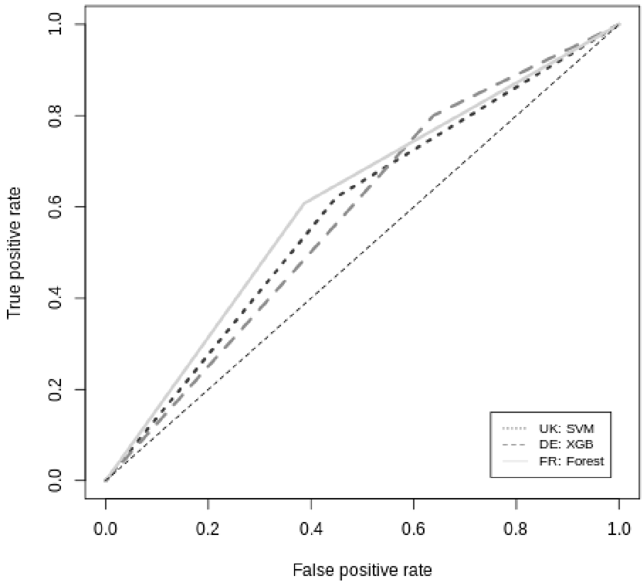

Figure 18: Model 1 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms (Pooled Data).

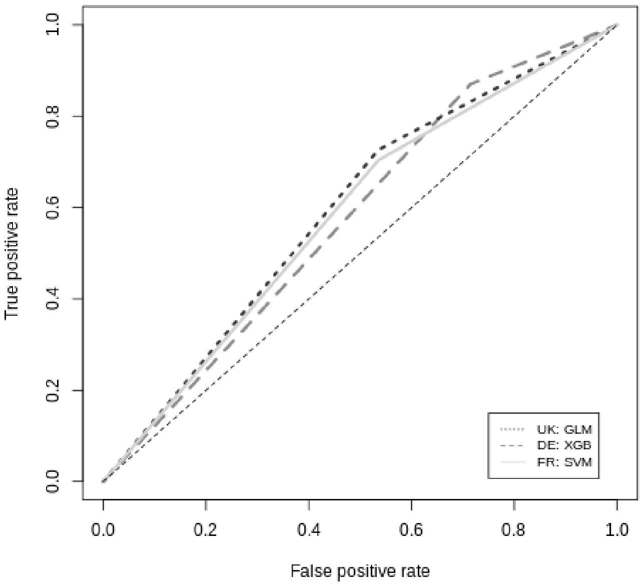

Figure 19: Model 2 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms (Pooled Data).

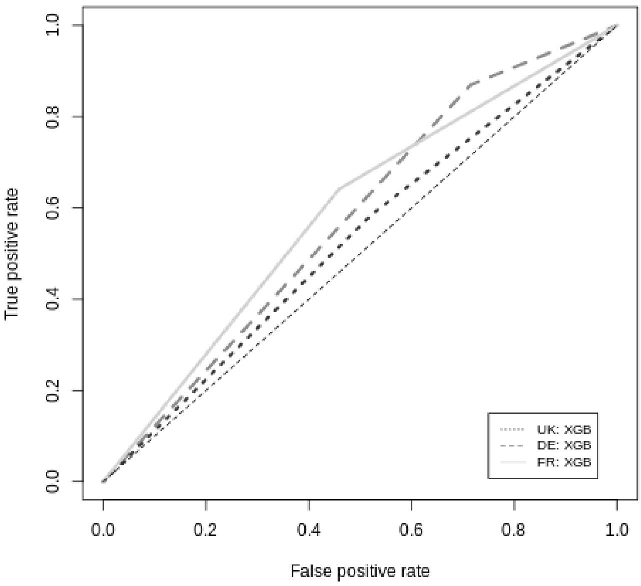

Figure 20: Model 3 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms (Pooled Data).

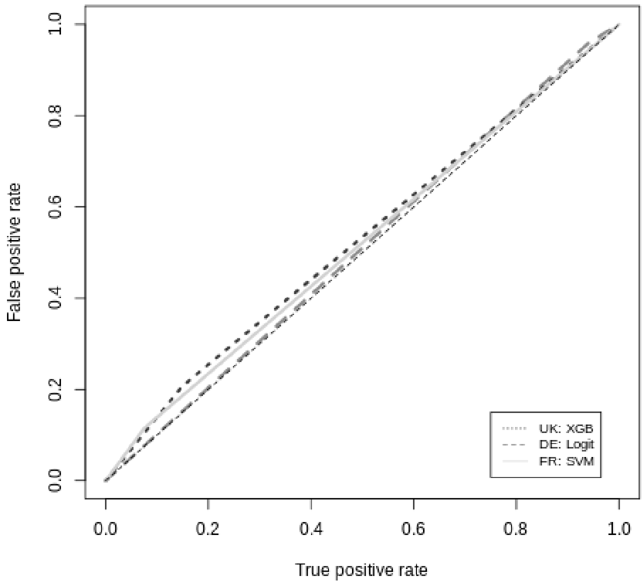

Figure 21: Model 4 ROC Curve Analysis Comparing the Prediction Power of the Competing Algorithms (Pooled Data).

Next, the results of the analysis of the five most important variables for the most successful algorithm per country are presented in Table 43 in descending order. Overall, it can be noted that different variables have a different importance for the predictions in the country. In the UK, the number of household members was the most important variable for the GLMNET to base the predictions on. In Germany and France, the most important variable was the completion rate. Being in paid work was only among the five most important variables in the UK, while the interviewer age was only among the five most important variables in France. The first contact hour was the fourth most important variable for the algorithms in both the UK and Germany. In the UK, the first contact workload was not among the five most important variables but was the second most important variable for the XGB prediction in Germany and the third important variable for the random forest in France. The age of the respondent as well as the years spent in education were among the five most important variables for all countries. In the UK and Germany, the respondent's age was the third most important variable for the GLMNET and XGB, respectively, while it was the second most important variable for the random forest in the data for France. The years spent in education are the fifth most important variable for all algorithms in the three countries.

United Kingdom	Germany	France
Model 2 – GLM	**Model 1 – XGB**	**Model 1 – Random Forest**
# Household members	Completion rate	Completion rate
Main activity: Paid work	First contact workload	Respondent age
Respondent age	Respondent age	First contact workload
First contact hour	First contact hour	Interviewer age
Education years	Education years	Education years

Table 43: Five Most Important Variables for Prediction for Best Performing Algorithms per Country, Descending Order (Pooled Data).

6.3. Discussion & Conclusion

Due to the large increase in observations, it is an expected finding that the Chi²-tests for the bivariate analyses in the pooled data are more likely to become significant. Even in cases where only the test result from one single round was significant, while relationships from other rounds were far from the 5% significance threshold, the overall test result for the pooled data almost becomes significant (see for example respondent's sex for the UK in Table 27). Due to the large sample size the relevance of bivariate associations should therefore not be overstated.

The findings on relationships over time demonstrate how important it is to measure concepts multiple times to avoid reading too much into chance patterns in single waves or rounds. The main takeaway message here is that the results from Chapter 4 might have changed noticeably for almost all variables and for all countries if a different ESS round was analysed. Research Question 1 (*'Are the associations between the predictor variables and the outcome stable over time?'*) must be answered with a decisive 'no'. The associations between the predictors and the outcome are predominantly volatile across time. That does not mean that the results from Chapter 4 are false. However, they are only reflecting the evidence based on this particular ESS round, which means that some of the relationships might change when analysing other rounds as the analysis in this section demonstrated. Only a minority of the findings and relationships proved to be constant over time. A danger that becomes apparent here lies in deriving recommendations from analyses that rely on a single measurement point. Recommendations based on an analysis from Round 8 for example, and jumping from correlation to causality, could result in the decision to only hire older interviewers, since successfully contacted units tended to be visited by older interviewers. If the recommendations were based on ESS Round 9, exactly the opposite recommendation would apply. The findings do not invalidate findings from Chapter 4. However, they illustrate that the findings in Chapter 4 need to be contextualised in their respective ESS round. Deviations in the findings are not necessarily bad or negative for the analysis. In fact, they can build a starting point for new analysis in order to investigate the causes for the changes in these relationships.

The literature review (see Section 2.2) brought to light several varied findings for the same variables either between countries or even within the same country. One possible explanation in previous chapters was that the concepts might not have followed the same operationalisation between the countries, i.e., have been measured differently. Another explanation theorised that the divergences could be explained by differences in survey implementation, population characteristics and cultural differences (Blom 2012). Based on the analysis of relationships over time in this chapter, it can now also be shown that the findings might not only vary between countries, but that they are also largely unstable over time and vary when measurements from different points in time are compared even within the same country. This supports Durrant and Steele's (2009) findings, who demonstrate an instability for some predictors, which change their influence on contact depending on which survey is being analysed even though all of the surveys were fielded in the UK. Blom's (2012) work in particular brought to light a variety of cross-country differences between the influences of variables on contact success: the author finds that the degree of urbanicity, the presence of access impediments, housing quality, interviewer gender, level of education, interviewer workload and call scheduling have different influences on contact success depending on the country under investigation. Although the directions of the relationships were not investigated in the analysis of this chapter, they do extend Blom's (2012) findings in a sense that the relationships not only seem to vary between the inspected countries but also that the meaningfulness of this relationship varies even within the countries depending on which ESS round is investigated. This means that any cross-country comparison of relationships between variables might not only be confounded by cross-country differences but also by time effects if surveys from different points in time are compared. It consequently remains debatable whether such comparisons are useful at all.

As desired, the pooling led to the expected overall increases in sample size. Over all countries even the model with the largest number of variables contains about six times as many observations in the pooled data compared to the single round dataset from Chapter 5. While Model 1 for the UK only featured 400 cases in the analysis for Chapter 5 the enlarged model can now be trained on 3,924 complete cases and even the smallest train model (2,022 cases, Germany, Model 1) consists of more cases that were available in all single models 3 of Chapter 5.

Despite these large increases in sample sizes Research Question 2 ('*Does the increased sample size lead to better predictions?'*) cannot be answered unambiguously. The results show that the increases in sample size led to large AUC gains for the larger Model 1. However, these gains dwindle when the number of predictor variables gets reduced despite an increase in the number of observations, to the extent that the algorithms based on the pooled data deliver poorer predictions than those algorithms that were trained using ESS Round 9 data only.

In Chapter 5 it was observed that there seem to be differences in the algorithms' computational costs with regards to the time they need for training depending on the number of observations they are trained on. It was shown that XGB deals particularly less well when applied to a limited sample size. While interesting patterns occurred in the analysis with the limited sample size, these patterns disappear when the extended data basis is used, as the ranking of the fastest algorithms almost unambiguously remains the same over all countries: in all countries and over all models, logistic regression was the fastest algorithm, followed by GLM. In all but two cases XGB came third, SVMs fourth and random forests were the slowest algorithm. The finding from Chapter 5, that XGB does well with increases in sample size, also finds support here.

The most important question remains whether the efforts to pool the data to generate gains in predictive power were worthwhile. While for Model 1 large gains in AUC values were obtained, these gains diminished in Model 2 and disappeared in Model 3 and 4. Meanwhile, computational costs for the training on the extended data basis increased tremendously. While the computation of all Model 1 of Chapter 5 took roughly 2.5 hours in total for all countries, all Model 1 calculations using the pooled data needed twice as long to compute. Thus, the average gain in AUC of 8 percentage-points over all countries was paid for in 5 hours of computation time. The efficiency gains for the most performant algorithms per country seem to be worth it: A comparison of the best performing algorithms per country of Chapter 5 (see Table 25) and Chapter 6 (see Table 42) shows that there is a difference in AUC of 6 percentage-points for the UK when utilising the pooled data compared to the ESS 9 data only. These gains were achieved by investing 2 more minutes in computational time. For both Germany and France, the best performing algorithms increased the AUC by 7 percentage-points. These gains came at the cost of 29 minutes and 60 minutes, respectively.

For Models 2 to 4 in all countries it remains debatable whether the gains in predictive power outweigh the computational costs. In Model 2 a 2 percentage-point increase can be observed in the best performing algorithm for the UK, which cost 12 minutes of computational time. While it is debatable whether the gains for the UK are worth the cost, the best performing algorithm for the German data, which used the pooled data, was only 0.1 percentage-point better than the best performing algorithm, which only used the ESS 9 data. Yet, this negligible gain was associated with a doubling in computational time. The French data is an extreme case of this as it shows, that the best performing algorithm using the extended data was even performing worse than the best performing algorithm only using ESS 9 data. In addition to that it also increased the computational time by 7 minutes. For UK Model 3 an AUC gain of 1 percentage-point can be observed, which can even be achieved using less computational resources, although the extended data was used. In Germany, the enriched data did not lead to a more successful model but instead resulted in an AUC loss of roughly 3 percentage-points, while costs remained roughly the same. In France, the AUC gain of 0.8 percentage-points was even achieved 20 minutes faster.

For Model 4 the gains for the UK disappear entirely, since the algorithms that utilise the extended data, do not outperform the predictive power of the algorithms that utilise the limited data. In Germany, an unsatisfying gain in AUC of 0.6 percentage-points can be observed, which was at least achieved at smaller costs. The French data shows a similar picture as the UK data, since the best performing algorithm using the extended data performs even worse than the best performing algorithm using the limited data, but at the cost of being roughly four times as time intensive.

Although it seems that some algorithms were able to gain increased AUC values while also reducing their computational time – which can be considered the best-case scenario – this might be a deceiving impression and must be discussed in more detail since there are some algorithms that simply take less time to be computed than others. The above-mentioned AUC gain of 0.6 percentage-points for Model 4 in Germany, for example, was obtained by a logistic regression, which was compared to an XGB. A logistic regression prediction without any machine learning pipeline and training process will always outperform the XGB in computational speed. Thus, a better approach might be to take a closer look at the average computation times per algorithm over all models and countries and compare these times to the computational times that were required for the analyses from Chapter 5.

Summed across all models and countries, random forests trained on the limited sample size of Chapter 5 needed 318 minutes to be trained, while they needed 2,089 minutes to be trained using the enhanced data of this chapter. The total computational time remained roughly the same for XGB (411 minutes versus 404 minutes) and logistic regression (0.01 minutes versus 0.04 minutes) but was noticeably higher for GLM (99 minutes versus 16 minutes) and SVM (1,634 minutes versus 1,020 minutes). Running all Chapter 5 algorithms of all countries uninterruptedly one after the other, would have taken a total time of 29 hours. Running all Chapter 6 algorithms of all countries uninterruptedly would take about 70 hours of computation time on the machine that was specified in Chapter 3.7.

Overall, it must be concluded that it is not worth investing this many computational resources for such small gains in predictive power at least for the countries under observation. Research Question 3 *('Do the gains in predictive power outweigh the expected increases in computational costs?'*) must therefore be answered with a clear 'no'. However, this does not mean that for other countries the same result has to be expected as the analyses have repeatedly shown the importance of country contexts.

This unexpected finding is interesting since it was anticipated that an increase in sample size would lead to better predictions. However, it seems that this is not necessarily true. The remaining question is why these increases in predictive power failed to materialise. One possible explanation might be that results might have looked different if a different set of algorithms was chosen. Yet, it seems unlikely that the absence of increases in predictive power can be attributed to a biased selection of algorithm, since at least XGB is widely considered one of the best performing algorithms developed so far.

The objective of this chapter was to find out whether increasing the sample size for the algorithms leads to gains in the predictive performance. While the associated analysis brought to light interesting insights for survey methodologists on the stability or instability of correlates of contacts over time, only modest performance gains were obtained despite all efforts. It was shown that only the models with more variables benefited from the larger sample size. However, the large sample sizes could not outweigh the lack in meaningful predictor variables of the smaller models. In addition, the algorithms for the smaller models not only contributed little predictive power, but also came with tremendously increased computational processing time. Therefore, the benefit of this approach remains questionable.

These findings do not necessarily imply that machine learning is not capable of predicting contact. It rather suggests that it is not feasible to efficiently predict first contact success based on a reduced variable set of paradata, even if a larger sample size is at hand. Instead, it suggests that a successful prediction depends on both a large sample size and meaningful predictor variables. It can be concluded that a prediction gets less valuable if either of the two is taken out of the equation. Another reason for the modest gains in predictive power, despite the pooling efforts, might be that the analysis was built upon a combined dataset consisting of nine independent survey rounds. These survey rounds cover a timespan of almost 20 years and the analysis showed that the included variable relationships might be too unstable to represent an unambiguous predictive relationship. Despite all efforts by the ESS researchers to make the fieldwork as comparable as possible, the rounds also varied a lot in their fieldwork characteristics for example because survey agencies who conducted the survey fieldwork changed over the years. This might further restrict their pooling suitability.

Despite these downsides and although the gains in predictive performance were smaller than expected, it was shown that predicting contact success is possible and increasing the sample size can be beneficial. In fact, the uncertainty of whether contact can be established or not can be reduced by approximately 7 percentage-points on average over all models and countries. If this reduction of uncertainty could be translated into real-world applications, fieldwork agencies would be able to allocate their resources more precisely.

For survey methodologists this chapter underlined and extended findings from other literature on contactability, while also promoting the use of machine learning algorithms for survey research. The analysis of the most important variables for the most successful algorithms shows that, for three of the countries it is important to have data on both the education and age of the respondent to successfully predict contact (see Table 42). If at least the data for the remaining four important variables was available for fieldwork planning, first contact success could be predicted in these countries already sufficiently well.

In this chapter an extended dataset was used, which combined all ESS Rounds 1 to 9 in order to maximise the sample size for the input models despite possible reductions due to listwise deletion. In a first step, the developments over time were investigated to find out which associations that were found in Chapter 4 are constant or vary between ESS Rounds 1 and 9. Despite the finding that the ESS questionnaire changed noticeably in the first five rounds, it must also be concluded that multiple associations are not stable over time. This finding is important when theorising about the potential influencing factors of contactability.

After the analysis of the associations over time, the models were fed to the same algorithms which had already been used in Chapter 5 to investigate whether the expected increases in predictive performance could be found. It turned out that an increase in sample size does not inevitably lead to performance gains. Instead, it appears that there seems to be a trade-off between a sufficiently large sample size but also having important variables in the model to base the predictions on. Thus, it turned out that not the model-algorithm combinations which had the highest number of observations were superior to all others, but instead it was the largest Model 1 in the case of Germany and France and the second largest Model 2 in the UK. These models still feature almost ten times the number of observations of their Chapter 5 counterparts. In a way it can thus be summarised that the main finding of this chapter was that it appears that machine learning algorithms are not only greedy when it comes to the sample size, as they tend to work better with larger samples, but also greedy with regard to the number of variables, since they produced the best predictions for the models with the highest number of variables.

More than anything this chapter showed that a contact prediction is possible and can even lead to correct predictions in 61% of the cases. The next chapter will make use of this finding by utilising the predictive power of the most successful algorithm in each country and simulating the best possible contact attempt for each individual unit.

7. Fits Like a Glove: Tailoring the First Contact Attempt for a Prototype of a ‘European Social Survey Fieldwork Optimisation Simulation’ (ESS-FOPSim)

The analyses from Chapter 6 showed that predicting first contact attempt success is not only possible in all selected countries but can also lead to improvements in correctly predicting contact of 11 percentage-points compared to a coin toss. Yet, until now the insights from these analyses have not been used for purposes other than to find the best performing algorithm per country. This chapter will utilise the best performing algorithm in each country from Chapter 6 and make it the engine in a simulation approach, which aims to optimise fieldwork procedures and will thus carry the proposed title ‘Fieldwork Optimisation Simulation’ (FOPSim). The objective is to find those features of a first contact attempt, which maximise the probability of first contact success given a set of an individual’s characteristics. Reminiscent of the ‘Tailored Design Method’ (Dillman et al. 2014) and ‘Tailored Fieldwork Design’ (Luiten and Schouten 2013) this set of optimised contact attempt features will be called ‘Tailored First Contact’, as the features for the first contact attempt are tailored to the specific personal characteristics of a unit to maximise the contact probability.

7.1. Research Questions

The feasibility of a 'Fieldwork Optimisation Simulation' will be investigated and a prototype of a tool will be developed. To accomplish this, the objective is specified by finding answers to three distinct research questions:

1. For a selection of units with a pre-defined set of personal characteristics: What are the characteristics of a tailored contact attempt with the highest probabilities of contact for these units?
2. For all units of the pooled dataset: How do the features of the observed first contact attempt differ from those of the simulated tailored contact attempt?
3. For all simulated possible contact attempts: What distinguishes the combinations that are predicted to result in non-contacts from those which are predicted to result in contacts?

The first research question aims to contribute to the understanding of how to best contact a specific person given their immutable personal traits. This is of particular interest for fieldwork agencies if they know specific characteristics about a potential respondent prior to making contact. Having this kind of information might occur, for example, in a panel survey environment or when inferences about a unit's personal traits can be drawn from a known unit's geographical location and/or because of population distributions.

The second research question gives more insight into how a tailored first contact attempt would have looked in comparison to the observed first contact attempt that actually happened during fieldwork. This question thus focuses on evaluating the differences between the observed and the tailored first contact attempt features.

The third research question investigates focusses on the simulated first contact attempt combinations and the differences between those that are predicted to result in a successful and those that are expected to result in an unsuccessful contact. The aim is to contribute to the insights for practitioners on the question of which features make first contact attempts successful. This approach might be considered a theoretical quasi-experiment answering the question 'what would have happened if…' since the outcome for the same respondent will be observed for all possible combinations of first contact attempt features.

Overall, the questions also help in contributing to the understanding of first contact success in the context of a cross-sectional survey, like the ESS. Besides other transferrable insights from the previous remarks, this approach might be of interest if projections about upcoming fieldwork outcomes need to be made based on experiences from the fieldwork of previous rounds. In other words, this approach, or an extension of it, might be a useful addition to the 'Fieldwork Monitoring System' that was used in the ESS Round 9 discussed in Chapter 1. Under the assumption that the survey population in the upcoming round does not differ significantly from that of the previous round, a fieldwork agency could estimate the best contact attempt combinations for the previous rounds and use these predictions as data-driven estimates for the upcoming round.

7.2. Chapter Specific Methods

The input variables that were used in most of the models from the earlier chapters can be categorised into two distinct groups: those that fieldwork agencies cannot influence, referred to as 'immutable traits', and those that can be influenced, referred to as 'alterable attributes'. Table 44 shows which of the variables fall into which category. For example, a fieldwork agency cannot influence whether a potential respondent lives in an area with a large amount of vandalism in the immediate vicinity. However, they can influence at which hour of the day to visit a unit in such an area.

Immutable traits	Alterable attributes
Type of house	Day of the first visit
Physical condition of building	Hour of the first contact attempt
Access impediments	Interviewer sex
Amount of litter in immediate vicinity	Interviewer first contact workload
Amount of vandalism in immediate vicinity	Interviewer success rate
Respondent's main activity in last seven days	Interviewer age
Urbanicity	
Respondent sex	
Number of household members	
Respondent's education years	
Respondent's age	
Respondent's marital status	
Total household income	
Whether or not children live at respondent's home	

Table 44: Categorisation of Input Variables into Immutable Traits and Alterable Attributes.

By using this categorisation, it is possible to investigate which combination of the alterable attributes leads to the highest predicted contact success probability for a given set of immutable traits, by simulating all contact attempts.

The success of a first contact attempt given the immutable traits and one combination of alterable attributes will be predicted by the model-algorithm combination which yielded the highest AUC value per country in Chapter 6. More specifically GLM-Model 2 (AUC = 59.6%) for the UK, XGB-Model 1 (AUC = 58.1%) for Germany and RF-Model 1 (AUC = 61.1%) for France will be used.

To investigate the tailored contact attempt for a given set of immutable traits and thus find an answer to the first research question, three example and hypothetical potential respondents are invented. The first hypothetical unit is a 22-year-old male student, who lives in a big city – referred to as the 'student' archetype. The second unit is a 39-year-old female in paid work, living with her children in a small town – referred to as the 'working mum' archetype. The third person is a 75-year-old retired male living in the countryside, referred to as the 'retiree' archetype. All units have more immutable traits, which are displayed in Table 45 and described in more detail below.

	Student archetype			*Working mum archetype*			*Retiree archetype*		
Immutable traits	**United Kingdom**	**Germany**	**France**	**United Kingdom**	**Germany**	**France**	**United Kingdom**	**Germany**	**France**
Type of house	Single Unit			Multi-Unit			Single unit		
Physical condition of building	Good or very good			Satisfactory			Good or very good		
Access impediments	Yes			No			No		
Amount of litter in immediate vicinity	None or almost none			Large or very large amount			Large or very large amount		
Amount of litter in immediate vicinity	None or almost none			None or almost none			Large or very large amount		
Respondent's main activity in last seven days	Education			Paid work			Retired		
Urbanicity	Big city			Town or small city			Country village		
Respondent sex	Male			Female			Male		
Number of household members	1			3			3		
Respondent's age	22			39			75		
Respondent's marital status	-	None of these (single)		-	Married		-	Separated/divorced	
Whether or not children live at respondent's home	-	No		-	Yes		-	No	
Respondent's education years	13	14	12	13	14	12	13	14	12
Total household income	-	5	3	-	5	3	-	5	3
Interviewer first contact workload	30	50	32	30	50	32	30	50	32
Interviewer completion rate	50%	36%	51%	50%	36%	51%	50%	36%	51%

Table 45: Immutable Traits for the Different Archetypes.

To a large extent, the units' immutable traits are homogeneous across countries. There are three immutable traits that are homogeneous within a country but vary between countries and these require further explanation. First, since different input variable models are used for the best performing algorithm in different countries, some variables are not used in all countries. More specifically, both Germany and France are best suited to the Model 1 input variable set, which included all variables as operationalised in Section 5.1, while the UK simulation uses the reduced Model 2 input, which did not include variables on a respondent's marital status, their household income decile or whether they had children living in their home. Consequently, the archetypes also lack the information for these variables in the UK. Second, while most units' immutable traits were held constant across the countries, the variables on years spent in education, household income, interviewer first contact workload, and interviewer completion rate were fixed to the within country mean. This means that the archetypes underwent a country-average education, have a country-average household income decile and were visited by a country-average successful interviewer who had a country-average workload. This approach was chosen to avoid problems due to differences in the distribution of these variables between countries: the average first contact workload for interviewers of one country could be outliers in another country. Speaking of which, interviewer first contact workload and interviewer completion rate were previously defined as alterable attributes in Table 44. However, due to their continuous value range (0 to 1 for the interviewer success rate and the full range of all workloads for the interviewer workload), they would add many possible permutations to the simulation, which would overload the already complex computations. The reasoning for this decision will be illustrated in the next paragraph.

For each given set of immutable attributes per archetype described above, all possible combinations of the alterable attributes are generated, which simulate first contact attempts under varying circumstances. The alterable attributes as well as their value range and the arising total number of combinations to predict can be seen in Table 46. Interviewer age was excluded from the Model 1 input variable set in Germany during the model pre-processing of Chapter 6 because of a high number of missing values. Consequently, it cannot be simulated here. Thus, the total number of first contact alterable attributes combinations to predict are lower in Germany (196) compared to the total number of combinations that need to be predicted in the UK (18,676) and France (18,032), respectively. As mentioned before, the workload and success rates are not simulated, but fixed at country mean values. It can be seen that the number of combinations is already high since the outcome for roughly 18,000 combinations needs to be calculated for each unit in the UK and France. If the interviewer success rate, even only with two decimal points, and the interviewer workload with an exemplary value range of 1 to 48 were added as alterable attributes this would have increased the number of combinations to predict to roughly 90 million for each unit in both countries. Assuming a sample size of only 2,000 per country this would amount to about 350 billion combinations to predict for the UK and France only. This would exceed the locally available computational power and thus underlines the decision to treat interviewer success rate and workload as fixed attributes.

Alterable attribute	Value range United Kingdom	Value range Germany	Value range France
Day of the first visit	1 to 7 (7)	1 to 7 (7)	1 to 7 (7)
Hour of the first contact attempt	0 to 22 (23)	8 to 21 (14)	0 to 22 (23)
Interviewer sex	Male, Female (2)	Male, Female (2)	Male, Female (2)
Interviewer age	23 to 80 (58)	-	20 to 75 (56)
Number of Combinations	**18,676**	**196**	**18,032**

Table 46: Alterable Attributes, Associated Ranges and Total Number of Combinations to Predict per Observation per Country.

For each unit and combination, the underlying algorithm predicts the probability of a first contact success. Subsequently, from all predictions, the combination of alterable attributes per unit with the highest estimated probability of success is selected, in order to establish what the tailored contact attempt would look like for a given set of immutable traits, and how this tailored first contact differs by country.

Even though the tailored first contact attempt is only simulated for three units per country to answer the first research question, this already implies a total number of 110,712 predictions across countries (56,028 predictions for the UK, 588 for Germany and 54,096 for France).

To answer the second research question, the approach had to be adjusted slightly. This aims to predict the tailored first contact attempt for all units of the pooled ESS dataset (see Section 3.4 and the analyses in Chapter 6) that were included in the respective model. Calculating every combination for all available units in the respective model was not possible because of insufficient locally available computational resources. To reduce the number of combinations, interviewer age was treated as a fixed attribute for the second research question. This reduces the number of combinations per unit to 322 in the UK and France. Overall, this leads to a reduction in the total number of combinations to predict to a total of 4,641,568 for the second research question. While this limits the available information for the tailored first contact attempt, it does at least show that the simulation can be adjusted to the needs and resources of a user.

7.3. Results

Table 47 shows the five combinations with the highest predicted probability of contact for each of the archetypes. One general finding for the UK is particularly striking: independently of the archetype and their immutable traits, the GLM-Model 2 always predicts a similar tailored contact attempt. Regardless of whether a student, working mum or retiree needs to be contacted, it is suggested to have first contact attempts on Sundays at 10pm with only slight variations in the interviewer's age between 76 and 79 years. Another striking finding is that in France, although more than 18,000 combinations were tested for the 'student', only two of these were predicted to yield a first contact success probability greater than 50%. While these two combinations are categorised to result in a contact, the associated contact probability is just slightly above 50%. Although these two combinations show that it is difficult to predict a contact success for the student archetype in France, the chances are best when the student is contacted on Wednesdays at 1pm by a female interviewer. Besides these eye-catching results, findings for Germany with regards to the student archetype are less conspicuous: In Germany, the most successful contact combinations for the student are on Wednesdays and Saturdays in the afternoon or in the evening respectively. Furthermore, four of the five best contact attempt combinations suggest a female interviewer. While the contact probabilities for the student are just above 50% in France, they range from 61.7% to 66.9% in Germany.

For the 'working mum' archetype, the random forest based on the Model 1 data for France predicted the best first contact combination to be on Wednesdays at 7pm by a female interviewer who is between 46 and 50 years old. Results for Germany based on the XGB-Model 1 algorithm were more diverse. Four of the five highest probabilities were estimated for Wednesday mornings between 8 am and 11am and suggested that the earlier morning contacts should be made by a male interviewer while the 11am contact attempt should be done by a female interviewer. Besides, Saturday afternoon visits by a male interviewer yielded the second highest contact success probability. Contact probabilities for the working mum archetype ranged from 66.7% to 67.6% in France and from 73.6% to 74.6% in Germany. Similar results can be found for the retiree archetype in France. Wednesday contact attempts between 12pm and 1pm by a 44 or 45-year-old female interviewer are predicted to result in contacts with the highest probability. In Germany on the other hand, Monday evening attempts are predicted to be the most successful in combination with male interviewers, while Tuesday evening attempts are predicted to be successful when female interviewers try to establish contact.

Contact probabilities for the working mum archetype ranged from 80.4% to 81.0% in France and from 80.7% to 83.1% in Germany. For each country the lowest contact probabilities can be found for the student, followed by the working mum archetype, while the highest contact probabilities can be found for the retiree archetype.

		United Kingdom		Germany		France	
		Probability	Combination	Probability	Combination	Probability	Combination
Student	1	52.70	Sun, 10pm, 80, male	66.91	Wed, 7pm, Female	50.15	Wed, 1pm, 36, Female
	2	52.69	Sun, 10pm, 79, male	66.45	Sat, 7pm, Female	50.07	Wed, 1pm, 37, Female
	3	52.68	Sun, 10pm, 78, male	62.44	Sat, 7pm, Male	-	-
	4	52.67	Sun, 10pm, 77, male	62.20	Wed, 4pm, Female	-	-
	5	52.66	Sun, 10pm, 76, male	61.71	Sat, 4pm Female	-	-
Working Mom	1	57.22	Sun, 10pm, 80, male	74.59	Wed, 11am, Female	67.66	Wed, 7pm, 49, Female
	2	57.21	Sun, 10pm, 79, male	73.96	Sat, 4pm, Male	67.45	Wed, 7pm, 48, Female
	3	57.20	Sun, 10pm, 78, male	73.70	Wed, 10am, Male	67.07	Wed, 7pm, 47, Female
	4	57.19	Sun, 10pm, 77, male	73.58	Wed, 8am, Male	66.75	Wed, 7pm, 46, Female
	5	57.19	Sun, 10pm, 76, male	73.58	Wed, 9am, Male	66.7	Wed, 7pm, 50, Female
Retiree	1	69.10	Sun, 10pm, 80, male	83.10	Mon, 8pm, Male	81.02	Wed, 12pm, 45, Female
	2	69.09	Sun, 10pm, 79, male	83.10	Mon, 9pm, Male	80.94	Wed, 1pm, 45 Female
	3	69.09	Sun, 10pm, 78, male	82.95	Tue, 8pm, Female	80.57	Wed, 2pm, 45, Female
	4	69.08	Sun, 10pm, 77, male	82.95	Tue, 9pm, Female	80.50	Wed, 12pm, 44, Female
	5	69.07	Sun, 10pm, 76, male	80.65	Mon, 7pm, Male	80.43	Wed, 1pm, 44 Female

Table 47: Top 5 Combinations with Highest Predicted Contact Probability by Country.

The results for the simulation of the tailored contact attempts for all units of the pooled dataset are presented next. Table 48 summarises the alterable attributes of the observed contact attempts and the tailored contact approach. For all available 9,423 units in the UK, 2,890 in Germany and 3,301 in France the tailored first contact attempt was predicted. In all countries a tailored first contact attempt would increase the percentage of contacts by a large proportion: while 53% of all units in the UK were observed as contacts, about 92% of the units are predicted to be contacts using a tailored first contact approach. Similar results can be found for Germany, where a tailored first contact attempt increases the share of first contacts to about 95% from 64%. While in France the increase in the proportion of contacts is not as high as for the other countries (50% versus 73%) it is still a meaningful improvement.

In Germany, 57% of the tailored first contact attempts feature a female interviewer, which diverges noticeably from the observed interviewer sex distribution of 37% female interviewers. In France, only 27% of tailored attempts involve a female interviewer compared to 35% observed female interviewers. For the UK, 100% of all optimal contact attempts featured a male interviewer, which diverges suspiciously from the observed data.

As with previous results from the example cases, the highest chances for contact were estimated for Sunday evening calls at 10pm for all 9,423 units in the UK sample, which is a large divergence from the observed data. For Germany, tailored first contact attempts would be most likely to happen on Saturdays, while Tuesdays, Mondays and Sundays also seem to be valuable alternatives. Wednesdays, Thursdays and Fridays were less commonly represented in the tailored first contact attempt combinations. The reduced importance of Wednesday, Thursday and Friday in the tailored attempts in Germany diverge noticeably from the observed contact attempts, which were almost evenly spread out over the week, except for Sunday attempts, which were less prominent. Interestingly, the tailored first contact attempts prioritise the weekend more than can be seen in the observed data: instead of the observed 15% of contacts on Saturdays, 38% of the tailored attempts would happen on a Saturday and another 12% on Sundays (compared to only 2% for observed Sunday attempts). In France almost half of all tailored contact attempts (48%) would happen on Wednesdays (in contrast to 14% of observed contacts), while almost another 30% of the predicted contact combinations suggest contact on Thursdays or Saturdays. This is particularly interesting because the observed data shows a clear preference for Saturday attempts. For both Germany and France, the average best predicted contact time lies between 3pm and 4pm, which is the same as the observed first contact hour.

	United Kingdom (n=9,423)		**Germany (n=2,890)**		**France (n=3,301)**	
	Best Predicted	**Observed**	**Best Predicted**	**Observed**	**Best Predicted**	**Observed**
% Contacts	92.41	53.00	94.60	63.53	73.43	50.08
% Non-Contacts	7.59	47.00	5.40	36.47	26.57	49.92
% contacts made by						
Female interviewer	0.00	56.46	56.82	36.85	27.29	34.69
Male interviewer	100.00	43.54	43.18	63.15	72.71	65.31
First contact attempt day in %						
Monday	0.00	18.44	15.74	16.12	9.21	14.45
Tuesday	0.00	22.03	17.27	18.17	5.70	15.75
Wednesday	0.00	18.35	6.06	15.74	48.38	13.51
Thursday	0.00	17.09	2.08	17.40	13.54	15.03
Friday	0.00	12.89	8.93	15.12	4.79	13.15
Saturday	0.00	8.11	37.51	15.43	15.96	27.81
Sunday	100.00	3.09	12.42	2.01	2.42	0.30
Average hour of first contact attempt (sd)	22 (0.00)	14.47 (2.91)	15.67 (3.82)	15.35 (2.72)	15.18 (4.80)	15 .09 (2.92)

Table 48: Characteristics of Best Prediction versus Observed Characteristics.

Figure 22 to Figure 24 show heatmaps of the frequency of the observed time and day combinations of the first contact attempts, plotted at the top of the figures, versus the time and day of tailored first contact attempts. For the UK, the heatmap shows all 9,423 tailored contacts to be on Sunday evenings at 10pm. In Germany, the most prominent time of day and day of week combination for the tailored attempts is Saturdays at 7pm. Generally, in Germany Saturdays as well as Sunday mornings and Monday and Tuesday evenings are common among the tailored contact attempts compared to the prominence of early week afternoon attempts in the observed data. In France, the most frequent time of day and day of week combination for the optimised contact attempt is Wednesdays at 1pm, compared to the dominance of Saturday contact attempts in the observed data. Wednesdays as well as Saturdays are prominent among the tailored first contact attempts at all times of the day, while for most other days only the evening hours are suggested.

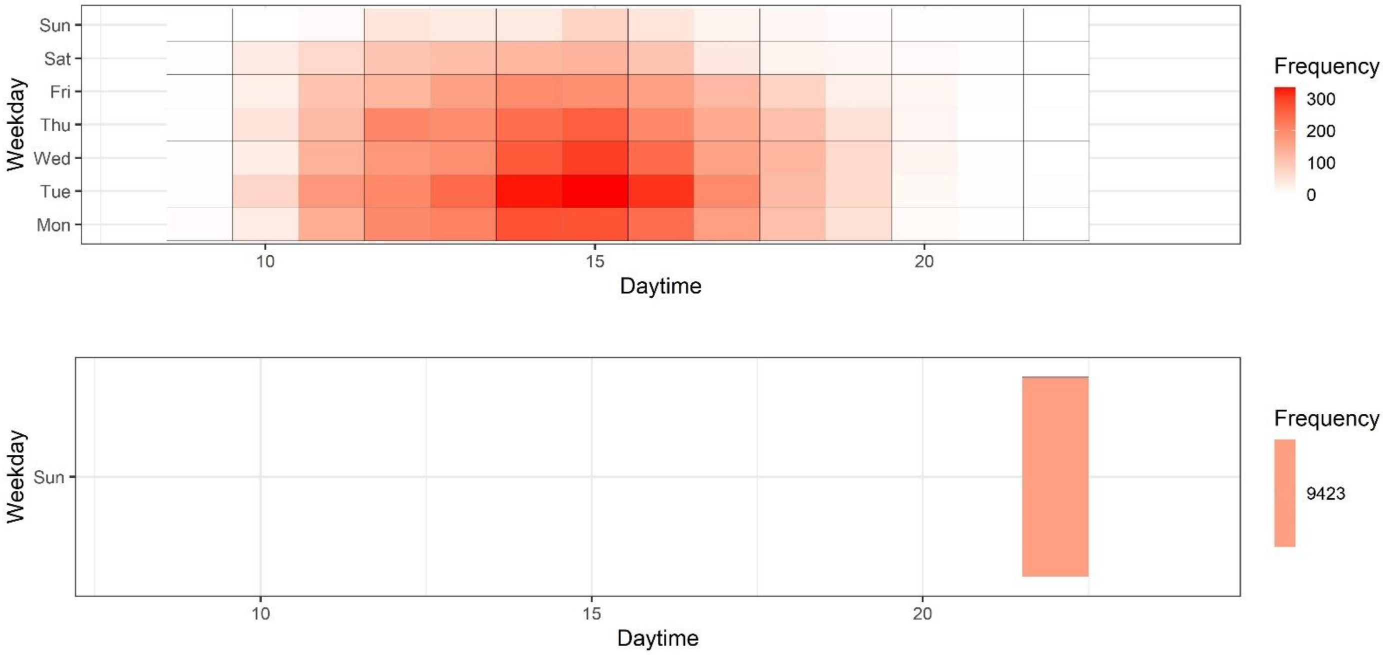

Figure 22: Heatmaps of Frequency of First Contact Attempts by Hour of the Day and Day of the Week in the UK. Observed (top) versus Optimal (bottom).

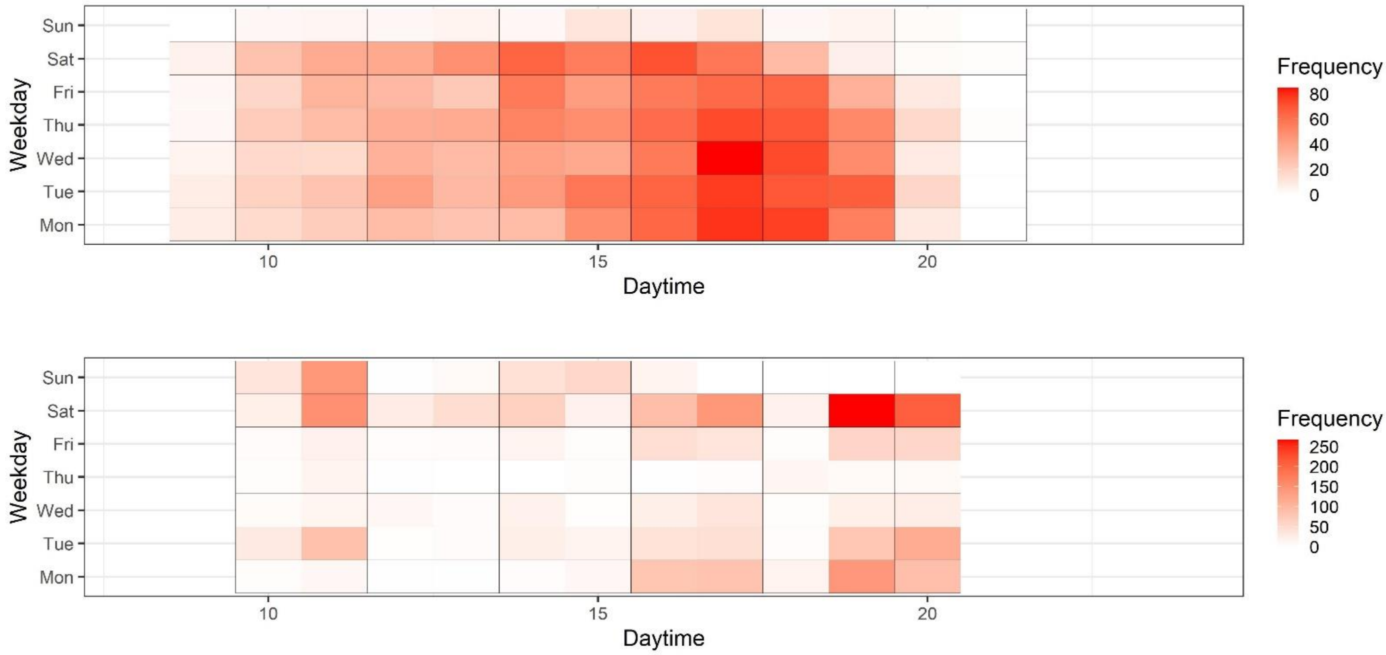

Figure 23: Heatmaps of Frequency of First Contact Attempts by Hour of the Day and Day of the Week in Germany. Observed (top) versus Optimal (bottom).

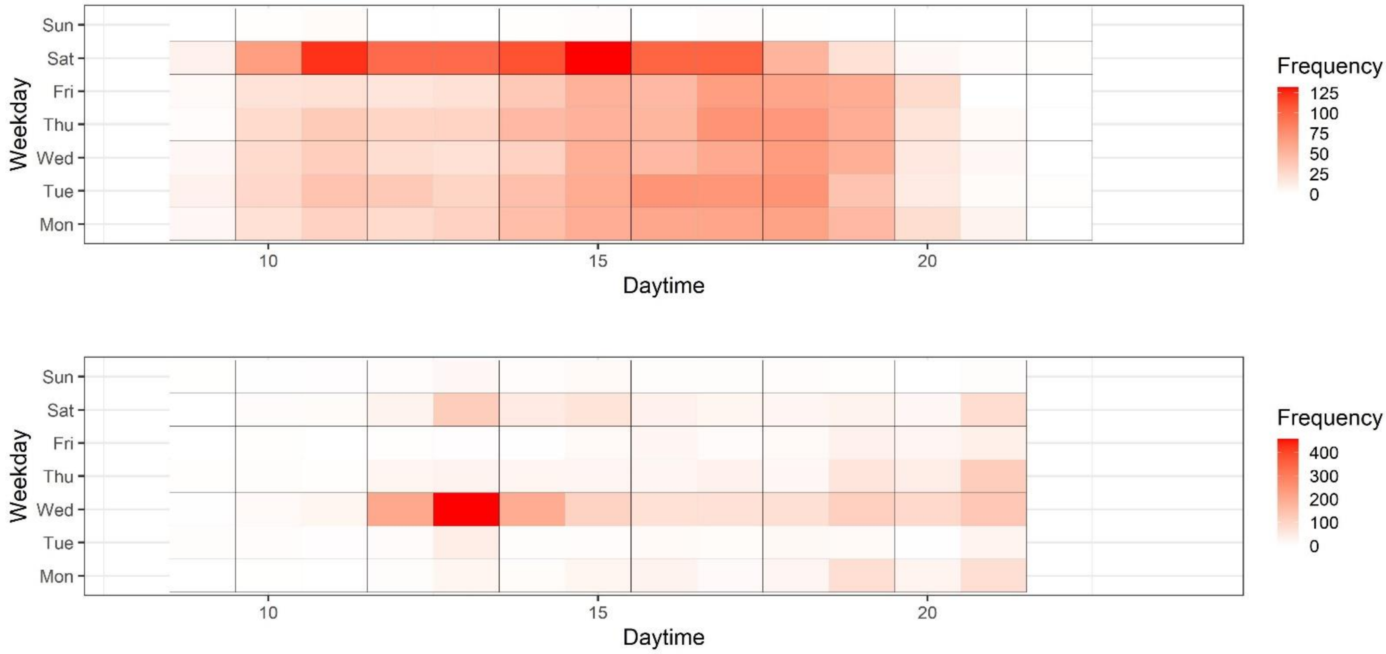

Figure 24: Heatmaps of Frequency of First Contact Attempts by Hour of the Day and Day of the Week in France. Observed (top) versus Optimal (bottom).

In the next paragraphs, the results of the investigation into what makes a successful tailored contact attempt, are presented. Table 49 shows the characteristics of all those combinations that are predicted to result in a successful contact, against those combinations that are predicted to result in a non-contact. Through this, the characteristics can be examined that make for a successful combination. In the UK, the combinations that are predicted to result in contacts predominantly feature male interviewers, which is different from Germany and France, in which successful combinations distinguish themselves by a larger share of female interviewers. While for the UK almost no differences can be seen in the day of the week between successful and unsuccessful simulated contact attempts, there is a large difference in the average time of day. While those combinations that are predicted to result in contacts have an average first contact hour at about 8am, those with a predicted non-contact have an average first contact hour of 3pm. This is particularly interesting since all the first contact combinations with the highest probability of contact unanimously suggested 10pm as the best contact time. In Germany, the average first contact hour hardly differs between the groups. However, the contact days are different. While in the group of predicted non-contacts 21% of these attempts are expected to happen on Sundays, there are only 11% of Sunday attempts in the predicted contact group. Similarly, while there are 11% of Friday contact attempts in the group of predicted non-contacts, there are 16% of Friday attempts in the group of predicted contacts. In France, there are only minor differences between the groups with regards to the day of the week. It is clear however, that in the group of predicted non-contacts there are only 12% of Wednesday attempts, while in the group of predicted contacts there are almost 17% of Wednesday attempts.

	United Kingdom		**Germany**		**France**	
	Predicted Non-contact (n=916,237)	Predicted Contact (n=2,117,969)	Predicted Non-contact (n=200,029)	Predicted Contact (n=366,411)	Predicted Non-Contact (n= 577,813)	Predicted Contact (n= 485,109)
% Contacts made by						
Female interviewer	51.66	48.10	49.28	50.39	46.13	54.61
Male interviewer	48.34	51.90	50.72	49.61	53.87	45.39
First contact attempt day in %						
Monday	14.32	14.24	12.75	15.12	14.85	13.62
Tuesday	14.91	13.57	12.49	15.27	15.03	13.40
Wednesday	14.39	14.16	14.06	14.41	12.33	16.61
Thursday	14.32	14.24	14.78	14.02	14.33	14.23
Friday	14.32	14.24	14.44	14.20	14.92	13.53
Saturday	14.22	14.36	10.88	16.15	14.40	14.14
Sunday	13.50	15.18	20.59	10.84	14.14	14.46
Average hour of first contact attempt (sd)	15.08 (5.13)	7.44 (5.67)	14.07 (4.17)	14.73 (3.93)	10.87 (6.70)	11.15 (6.5)

Table 49: Differences in Characteristics Between All Predicted Non-Contact Combinations and All Predicted Contact Combinations by Country.

7.4. Discussion & Conclusion

The analysis has shown that predicting optimal contact attempt combinations is possible. Yet, the results need to be discussed from a variety of perspectives.

First, it can be concluded that the predictions for the archetypes showed some expected results. The immutable traits of the 'student' archetype, for example, were deliberately chosen to establish an archetype of a hard-to-contact unit. The literature review in Chapter 2 as well as the analyses in Chapter 4 suggest that multiple of the characteristics of the student archetype (living alone, young, living in a big city, having access impediments) hamper its contactability.

Thus, it is not surprising that in France only two out of all combinations were predicted to result in a contact and that those only have a predicted first contact probability of just above 50%, which indicates the uncertainty of the prediction. While the most optimal contact days differ between France and Germany, it appears that contacting young male students, who live alone, is most successful when female interviewers are reaching out to them.

The 'working mum' archetype deliberately features immutable traits that should make this archetype easier to contact than the 'student' (more household members, children at home), but combines this with some other traits, which are expected to hamper the contact (being in paid work). Consequently, the algorithms are more certain in predicting contact compared to the student archetype (as indicated by higher probabilities in Table 47) but less certain in their prediction compared to the 'retiree' archetype. While Wednesdays appear to be beneficial for making contact with this archetype in both France and Germany, the optimal contact hours differ between the countries: while evening contacts are preferable in France, the results show that morning hours have the highest contact probability for this archetype in Germany. Additionally, female interviewers are preferable for this archetype in France, while four of the five best combinations for this archetype in Germany featured male interviewers.

The 'retiree' archetype was designed to feature those immutable traits that should make contact easy (older age, retired, living in the countryside) and in fact the algorithms predicted the contact with high probability values for this archetype in all three countries. There are, however, large differences between the countries in the combinations of the alterable attributes for this archetype. While in France Wednesday contacts around noon by female interviewers showed the highest probability, there were multiple contact days in Germany that were associated with a high probability, and evening hours were preferred overall while both female and male interviewers are predicted to be successful in establishing contact depending on the other features. The variety of different possible contact attempt combinations for Germany show that this archetype is easier to contact and that the best approach is not limited to a very narrow choice of combinations. This is important if a large number of contact attempts need to be planned using limited resources, as explained later.

The results for the UK were unexpected and can probably be attributed to one of three explanations: there is a chance that the random forest algorithm might not have worked as intended. This would certainly be a surprise since the analyses of both Chapter 5 and Chapter 6 were unobtrusive. Secondly, the initial predictions in Chapter 6 could have been influenced by very important outliers with high leveraging potential. If the ESS data included multiple contact attempts on Sundays at 10pm by older male interviewers, which resulted in contacts, because the attempt was planned, this might carry the risk of skewing the predictions in this chapter and all other predictions. Outliers were not deleted from the data in the previous chapters when the algorithms were trained. It was deliberately refrained from deleting outliers to process the data as completely and naturally as possible. Yet, it is unlikely that this is explaining the observed effects for the UK predictions in this chapter: out of all 9,423 observations which were included in the UK Model 2, there was only exactly one contact attempt that happened at 10pm. On top of that, while this single contact attempt did in fact result in a success, it did not happen on a Sunday but on a Tuesday. Additionally, of all Sunday attempts (291 of the 9,423) only four happened after 8pm and none at or after 10pm. Consequently, it is unlikely that outliers skewed the predictions. Lastly, it might just be possible that Sunday evening attempts by male interviewers in their late 70s are in fact extraordinarily effective based on the data that was fed into the algorithms. Unfortunately, only an empirical study or experiment can prove or refute this hypothesis. Out of all these three explanations, the third is the only one that cannot be rejected and is thus accepted until its rejection.

The differences between the characteristics of the tailored first contact attempt and the observed first contact attributes are insightful for survey methodologists. While in Germany the units which were included in Model 1 were primarily visited by male interviewers, the optimisation suggests appointing more female interviewers to maximise contact success probability. In France, on the other hand, there were three times more tailored combinations which featured male interviewers than female interviewers. Based on these results, it appears that fielding female interviewers would be more successful for first contact attempt success in Germany than in France.

The results for the predicted 'best day of contact' show large differences for the observed day of contact. While in Germany all days of the week had a similar proportion of contacts, (with the exception of Sundays), almost half of the optimal contact predictions featured weekend contact attempts and another 32% on Mondays or Tuesdays. Based on this data, it appears that switching from weekday to weekend contact attempts might be an effective way of maximising first contact attempt rates in Germany. Results for France, on the other hand, suggest the opposite. While in reality, the fieldwork focussed a lot on Saturdays for first contact attempts, the results from the prediction suggest reducing the number of Saturday attempts and instead trying to reach out to potential respondents more often on Wednesdays.

The findings for the time and day combination of the tailored contact attempts are mostly in line with the literature and previous analyses. For Germany the simulated optimal contact attempts should happen in the afternoon and early evenings or at weekends, as suggested in the literature (Campanelli et al. 1997; Durrant et al. 2011; Durrant and Steele 2009; Lipps and Benson 2005; Purdon et al. 1999; Stoop 2005, p. 160f; Vicente 2017; Wagner 2013; Wang et al. 2005; Weeks et al. 1987). Interestingly, the pattern is slightly different for Tuesday mornings, which are also predicted to result in successful first contact attempts. While these findings are also true for France, the predictions support this less well. There is indeed a high frequency of simulated contact attempts in the afternoon and early evening hours, but there is also a large proportion of early day contact attempts, particularly on Wednesdays. It seems that the predicted optimal time disagrees with the literature.

From the analyses of unsuccessful and successful alterable attribute combinations it can be seen that earlier contact hours might be beneficial for the UK, while unsuccessful contacts in Germany predominantly happen on Sundays, indicating that Sundays should not be overvalued. The main difference between unsuccessful and successful simulated combinations in France comes predominantly from the interviewer sex distribution suggesting that female interviewers lead to more successful first contact attempts compared to male interviewers.

Overall, it can be stated that if a fieldwork agency wanted to tailor the first contact attempt to the traits of a unit, they could feed the immutable attributes of this unit to an algorithm and get back the tailored first contact attempt with the highest predicted contact success probability in return. If fieldwork agencies do not know the immutable attributes of a unit before reaching out to them, they could still at least determine the distributions of the parameters of the tailored first contact attempt and plan their fieldwork accordingly. For example, a fieldwork agency planning fieldwork for ESS Round 10 could feed all observations of Round 9 to the simulation as the testset, as in the analysis above, and see the which features were predicted to be beneficial for first contact. The agency could then decide whether they want to prioritise those features the simulation deems important. In both cases, knowing traits beforehand or not, this approach can possibly serve as the starting point to plan fieldwork operations more effectively or at least contribute to the existing information.

However, there are three important constraints that need to be discussed. First, it needs to be remembered that even though a tailored first contact attempt with the highest probability can be selected, the underlying algorithm still only has a given performance as explained in Chapter 6. More precisely, this means that in the case of the UK simulation, which relies on the GLM-Model 2, almost 40% of the predictions can still be wrong since the underlying algorithm only had an AUC of 59.6. Similarly, results might be wrong in 42% considering the underlying AUC of 58% in Germany and 39% considering the AUC of 61% in France, respectively. Whether this makes this simulation approach practicable or not, needs to be tested in a fieldwork experiment.

Second, a valuable further feature to implement in such a tool would be to integrate an optimisation logic. In the examples used in this chapter, the contact attempt with the highest probability for each unit was selected in all cases regardless of the consequences and whether this might mean, for example, that all contacts need to happen at one specific point in time. An intelligent optimisation would factor in four important aspects: First, conditions set by fieldwork agencies for times of the day in which contact attempts are not possible would be considered and combinations would automatically be excluded if they fall into these prohibited time windows even if they yield high probabilities (for example Sundays at 10pm). Second, sparse resources would be considered. Even though technically the interviewer attributes are alterable, an agency might only have a very specific pool of interviewers with given traits. This means that there is only a fixed number of interviewers and only a specific age and sex distribution. A much larger team of interviewers would be required if all contact attempts were to be made at the same time of the day. Third, an intelligent optimisation technique would allow altering methodological restrictions: as the ESS quality guidelines suggest, an interviewer must not have more than 48 assigned interviews. Fourth, usually an interviewer also needs to be assigned to units in close geographical proximity to avoid long travel times. An intelligent optimisation method would thus consider and allow altering methodological restrictions while also optimising route planning when trying to find the tailored first contact attempt with the highest probability of contact. Dealing with restrictions and a limited budget also means that a more realistic prioritisation of contact attempt combinations is desirable. Until now only the contact success for a single unit is maximised even if maximising the overall first contact success for the complete sample is preferable. The simulation needs to be able to decide when to prioritise another sub-optimal combination, to manage the limited resources and simultaneously increase overall first contact success. Subsequently, scenarios could be generated by altering the limitations and restrictions and further insights could be generated and observed that might be helpful for fieldwork planning. One example might be to investigate the effect on first contact success when the allowed maximal workload per interviewer is doubled.

Future iterations of this prototype should consider including more immutable traits and alterable attributes as well as the addressed limitations and restrictions. If more attributes are included and their full range of values used it is recommended to make use of parallel computation techniques and/or switch to cloud computing instead of trying to run these operations locally.

The purpose of this last investigation of this thesis was to expand the findings from the analyses of Chapter 6 and to leverage the predictive performance of the most successful algorithms in each country by making them key components inside a simulation. For each individual and their unique set of immutable traits (like having children or not), this approach simulated all possible combinations of the alterable attributes of potential first contacts (like the time of the day). This approach did not only show that for each set of immutable traits a tailored first contact attempt could be predicted, but also allowed to investigate the differences in characteristics of successful and unsuccessful first contact attempts. Ideas for further extensions of the developed concept and prototype of the suggested Fieldwork Optimisation Simulation were presented and it was argued that such an approach could enable researchers to investigate first contact attempts in quasi-experiments. Moreover, the predictions that can be produced with the FOPSim could be used in real-world fieldwork procedures to evaluate whether they contribute to reducing survey fieldwork costs.

The final chapter will summarise the narrative and findings of this thesis, put them into perspective and set out some recommendations for future fieldwork operations.

8. Wrapping It Up: Survey Fieldwork Remains a Contact Sport

When a survey is in the field, researchers and practitioners face a difficult challenge: from a methodological point of view, it is necessary to establish contact with a sampled unit no matter how many contact attempts are required, to avoid the dangers of nonresponse and sample selection biases. Unfortunately, repeated contact attempts can come with high costs for the survey agencies as they invariably need to pay interviewers to conduct these contact attempts. Reducing the necessary number of contact attempts as much as possible is therefore in the interests of everyone who is involved in the survey fieldwork process. In a best-case scenario contact would be established at the first attempt for each unit. However, depending on their personal characteristics, units have different at-home patterns and, thus, their contact probability can vary considerably from unit to unit even at a given point in time. This thesis aimed to contribute to addressing this challenge, by investigating the feasibility of predicting the success of the very first contact attempt for three countries in the European Social Survey. In a first step, the topic was framed as a methodological problem and the importance of multiple contact attempts where necessary, to counteract sample selection and nonresponse biases was explained. Additionally, the extensive literature was summarised. It was shown that while research on contactability has been carried out for at least 30 years, the first contact attempt in the ESS has never been the primary focus of any research so far. The findings from the literature review enabled the identification of multiple important concepts, which correlate with contact success, for example the time of day a contact attempt is made. These concepts were then operationalised using data (respondent data and paradata) from the European Social Survey.

In Chapter 4 the first contact success in ESS Round 9 of the United Kingdom, Germany and France was extensively analysed by finding answers to four research questions. Despite large country differences, this analysis supported many findings from the literature review, including the positive effect of later hour contact attempts or negative effect of urbanicity on contactability. Even the large country differences themselves find support in previous research. However, it was also shown that even more research is needed since disentangling the various correlates from one another remains an obstacle.

Chapter 5 went beyond the bivariate analysis of the first contact attempts success and introduced a machine learning approach to predict the contact success of units in the European Social Survey Round 9 based on the variables operationalised from the literature. This implies the assumption of causality between the predictors and the target variable as part of the model fitting, which cannot be verified. The results show that there is not one single algorithm or input dataset that outperforms all others. The predictive performance varies considerably both within the countries as well as between them. Overall, a prediction is feasible in most cases. Yet, it must be noted that the predictive performance is not convincing, since some algorithms do not perform better than a random guess and even the best performing algorithms achieve a maximum AUC value of 61%.

One reason for the underwhelming performance of the algorithms in Chapter 5 was thought to lie in the limited sample sizes, which is why it was extended in Chapter 6 by pooling data from all available ESS rounds 1 to 9. The objective was to investigate whether the predictive performance of the deployed algorithms improves with increases in sample size. These increases in sample size are desirable as machine learning algorithms tend to work better with a higher number of observations. However, since most algorithms also only work well with complete cases, the risk of reductions in sample size (because of listwise deletion) increases the more variables are included in a model. Thus, the pooling was aimed to counteract any potential reductions in sample size because of listwise deletion, while also allowing for inclusion of as many variables as needed. In fact, it was observed that the overall performance of all algorithms improved. However, while it appears that the baseline was raised and all algorithms tend to have a better predictive performance, the headline of the predictions remained the same, which means that still the best performing prediction had an AUC of 61%. Yet, it was shown that the best performing algorithms can reduce the uncertainty of establishing contact by roughly 7 percentage-points on average.

The insights from the previous chapters culminate in Chapter 7, which applied the findings and introduced a proposal for a 'Fieldwork Optimisation Simulation' to tailor the first contact attempt to the individual traits of a target unit to maximise their contact probability. The chapter proved the feasibility of predicting the optimal contact attempt characteristics given a target unit's set of immutable character traits like their occupational status. However, the chapter also emphasised important extensions to increase the prototype's practicability. These extensions include considerations of prohibiting unethical fieldwork hours, considering scarce interviewer resources, allowing for modification of methodological restrictions as well as introducing and optimising an interviewer route planning system. Overall, it is suggested to enhance the prototype by an optimisation logic that accounts for these considerations. Such an extended prototype would maximise the first contact attempt for each target unit, while simultaneously controlling the resource budget. Unfortunately, testing the tool's practicability under real-world fieldwork circumstances was out of the scope of this research.

The findings from this thesis offer some recommendations and insights for both survey methodologists and practitioners.

First and foremost, it is clear that using machine learning approaches and simulations to answer survey methods research questions has proven to be informative. Even though today social scientists might not receive formal training in data science methods, it was demonstrated in the methods section (Section 3.8 and Section 3.9) of this thesis that the differences between the fields are not as large as they might be perceived. Social scientists should therefore feel confident to explore other areas of research where machine learning techniques can accompany traditional analyses. The suggested 'Fieldwork Optimisation Simulation' has shown its potential to tailor first contact attempts to the immutable traits of target units to maximise their first contact attempt contact probability. Further research could increase the efforts to extend the practicability of such a tool. Despite this important finding and the overall success of machine learning algorithms in predicting first contact success, it is important to highlight that – against all expectations – logistic regression models, even without any particularly complex specification, do not perform considerably worse than the more sophisticated machine learning algorithms in multiple cases. In fact, logistic regression predictions were better than those from complex models more than once. This raises the question for future research on how these logistic regression models would perform if they were specified with more complexity by adding interactions or multi-level structures for example. For practitioners who need to decide between running machine learning or logistic regression predictions, it might be helpful to consider their current capabilities: if the knowledge and staff for machine learning predictions is available, a machine learning approach could be pursued. If, however, neither knowledge nor staff or time is available, the investments in staffing and/or training might not outweigh the benefits of the machine learning prediction and logistic regression predictions could be used with a reasonable degree of confidence.

Second, the benefit of using contact protocols cannot be overestimated. All analyses in this thesis were only possible because fieldwork agencies, which carry out the ESS fieldwork, are stipulated to deliver detailed contact protocols for each contact attempt. The higher costs that arise from obliging interviewers to complete the protocols are offset by the value this information can create in analyses. Therefore, surveys where the research team are concerned to investigate all sorts of potential influences on survey quality should implement the use of contact protocols. Although the contact protocols already feature about 20 different questions on both the contact attempt itself as well as the information on the housing unit and area the potential respondent lives in, having even more information from especially the interviewers themselves would be beneficial for analyses of contact behaviour due to the interviewers' crucial role in the fieldwork process. The interviewers' amount of experience in conducting fieldwork is one example for a variable that might be of particular interest in future analyses since it can be theorised that more experienced interviewers also have more successful strategies in contacting and finally convincing target units. Experience could for example be measured by asking whether this is the interviewer's first contract with a fieldwork agency or how many survey fieldwork projects the interviewer has already worked on as an interviewer.

Additionally, more information about the interviewers' contractual status might prove useful in an investigation on contact success. It can be theorised that interviewer motivation is dependent on their contract type and/or their pay rate. If this information was collected one could investigate whether first contact attempts from interviewers on casual contracts are different from those of permanently employed interviewers, or whether there is a difference between interviewers that are paid on hourly base or per case. In the context of a comparative survey this is particularly interesting because interviewer working conditions can be seen as plausible sources of between-country contact variance and thus can impact data quality. Paradata like this should then be exploited to contribute to an even better understanding of the survey process and to control for country variance. Moreover, this also enables researchers to think critically about current survey recommendations and practices. In the case of the ESS, one could think of re-evaluating whether the strict guidelines for establishing contact (e.g., strict number of visits at certain days), which are equally mandatory for all countries, are even similarly effective in all countries. Maybe paradata shows, that these guidelines should be altered, loosened, or tightened from country to country. This could be proposed to the national coordinators to support their fieldwork efforts.

Third, most findings vary considerably across countries and time. Therefore, recommendations are at least country specific. The analyses from Chapter 4, which focus on just one ESS round, indicate that fieldwork procedures in Germany and the UK are largely in line with the literature's recommendations while fieldwork in France deviates noticeably from these. However, the analyses of multiple rounds show that – ceteris paribus – there are only few constant findings: fieldwork administrators in the UK can expect a constant relationship between contact success and the main activity in the past 7 days of a unit as well as whether their houses have access impediments. Furthermore, it appears that later contact hours are more favourable to a successful contact. Contacted units were approached by interviewers who are more successful in converting their workload into completed interviews, highlighting the importance of these interviewers. Fieldwork managers can expect to reach out more easily to older units with less formal education and who live in larger households, as the results for these relationships and differences were constant over all ESS rounds in the UK.

German survey fieldwork researchers can expect a consistent relationship between contact success and urbanicity as well as the type of housing unit. Similarly to the UK, the results suggest the benefit of later contact hours by more successful interviewers. Like in the UK, older units with less formal education living in larger households should be easier to contact in Germany on the first contact attempt.

Fieldwork administrators in France can expect a constant relationship between a unit's contact success and their type of housing unit, whether it has access impediments and the unit's main activity in the past 7 days. In contrast to the findings from the UK and Germany, it appears that earlier contact hours are positively associated with first contact attempt success. Therefore, French fieldwork managers might want to continue reaching out to units at earlier hours of the day and hire interviewers that proved to be more successful in finishing interviews. Besides this, units in France also appear to be more easily contactable when they are older, received less formal education and come from larger households. While these results support previous findings from the literature, the outcomes for the remaining variables varied too much over time, to derive any practical fieldwork recommendations.

Besides important findings and valuable recommendations this thesis also has three important limitations. Most importantly a lot of the analyses to investigate or predict the contact success of a unit included information for the unit which is only available from an interview. It is impossible to predict a unit's contact success based on whether they have children if the unit was never surveyed before. The approach in this thesis was chosen anyway for five reasons. First, in the context of the discussed Fieldwork Monitoring System, which was used in the ESS Round 9 and introduced in Chapter 1, estimates for the planning of future survey rounds, like fieldwork outcomes and cost, are based on information from previous rounds, for which this information is available. The analyses and prototype could contribute to this Fieldwork Monitoring System. Second, the analysis aimed to prove the concept of a predictive modelling approach and thus simply accepted this constraint. Third, in the initial phase of this dissertation project it was planned to successively replace the personal information by geo-spatial information from the sampling point. Unfortunately, such information was unavailable. Future research might be able to investigate whether replacing geo-spatial information as proxies for the personal information proves to be useful. Fourth, maybe sampling frames get extended in the future to contain more information on the sampled unit. This could, for example, be done by linking social media data to sampling units of a population register frame. Lastly, although this thesis dealt with cross-sectional data, the results might provide useful in the context of panel surveys in which information on respondents is available from previous rounds. This information can be exploited to predict the tailored contact attempt for the same unit in future panel waves.

Although the selection of the three example countries was particularly interesting because all of them suffer from high nonresponse, it would be interesting to find out whether predicting contact success is also possible in the remaining participating countries of the European Social Survey.

Lastly, some analysis pushed the locally available computation resources to their limits. Future research dealing with similar complex predictions should utilise both parallel computing and cloud computing resources, to speed up the computational time and avoid storage overflows.

Besides the already mentioned future research options, this thesis has paved the way for multiple further research possibilities. One interesting question that remains is whether there is a ‘Goldilocks Zone’ between number of variables and number of observations that maximises the predictive performance of the algorithms. Other options include a translation of this approach to telephone surveys or extending it to subsequent contact attempts and maximising not only the first contact attempt probability but the overall contact success and thus contributing to the ESS scientific standard of maintaining a 3% non-contact rate. The most exciting investigations might look specifically at the prototype again to investigate what the best contact attempts look like for specific subpopulations such as minority groups or find out whether the extended prototype can prove useful in a field experiment. This in fact might be particularly important to investigate what the 11 percentage-point increase in first contact success prediction translates to in terms of savings in fieldwork costs.

More widely, this thesis contributed to a better understanding of the first contact attempt in face-to-face surveys by investigating the ESS and suggested ways to improve this crucial phase during the fieldwork period. It laid the foundation for future research and advancements of the suggested prototype which carry the potential to greatly reduce the fieldwork costs of face-to-face surveys.

References

Aizerman, M., Braverman, E., and Rozonoer, L. (1964), "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, 821–837.

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017), "Explaining Recurrent Neural Network Predictions in Sentiment Analysis," *arXiv:1706.07206 [cs, stat]*.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011), "Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations," *International Journal of Methods in Psychiatric Research*, 20, 40–49. https://doi.org/10.1002/mpr.329.

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., and Heinemann, J. (2019), "Predicting Voting Behavior Using Digital Trace Data," *Social Science Computer Review*, 089443931988289. https://doi.org/10.1177/0894439319882896.

Best, H., and Wolf, C. (2010), "Logistische Regression," in *Handbuch der sozialwissenschaftlichen Datenanalyse*, eds. C. Wolf and H. Best, Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 827–854. https://doi.org/10.1007/978-3-531-92038-2_31.

Bethlehem, J. G., Cobben, F., and Schouten, B. (2011), *Handbook of nonresponse in household surveys*, Wiley series in survey methodology, Hoboken, N.J: John Wiley & Sons.

Biemer, P., and Lyberg, L. (2003), *Introduction to survey quality*, Wiley series in survey methodology, Hoboken, NJ: Wiley.

Biemer, P. P. (2010), "Total Survey Error: Design, Implementation, and Evaluation," *Public Opinion Quarterly*, 74, 817–848. https://doi.org/10.1093/poq/nfq058.

Blakely, E. J. (Edward J., and Snyder, M. G. (1997), *Fortress America : gated communities in the United States*, Washington, D.C. : Brookings Institution Press.

Blom, A. G. (2012), "Explaining cross-country differences in survey contact rates: application of decomposition methods: Cross-country Differences in Survey Contact Rates," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 217–242. https://doi.org/10.1111/j.1467-985X.2011.01006.x.

Botman, S., L., and Thornberry, O., T. (1992), "Survey Design Features Correlates of Nonresponse," in *JSM Proceedings*, Proceedings of the Survey Research Methods Section, Alexandria, VA: American Statistical Association, pp. 309–314.

Breiman, L. (ed.) (1984), *Classification and regression trees*, Boca Raton: Chapman & Hall [u.a.].

Brick, M. J., Allen, B., Cunningham, P., and Maklan, D. (1996), "Outcomes of a Calling Protocol in a Telephone Survey," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 142–149.

Brown, R. V. (1967), "Evaluation of Total Survey Error," *Journal of Marketing Research*, 4, 117–127. https://doi.org/10.1177/002224376700400201.

Bullington, J., Endres, I., and Rahman, M. A. (2007), "Open-Ended Question Classification Using Support Vector Machines," Carrolton.

Buskirk, T. D. (2018), "Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research," *Survey Practice*, 11, 1–13. https://doi.org/10.29115/SP-2018-0003.

Campanelli, P., Sturgis, P., and Purdon, S. (1997), "Can you hear me knocking? and investigation into the impact of interviewers on survey response rates."

Campbell, M., Hoane, A. J., and Hsu, F. (2002), "Deep Blue," *Artificial Intelligence*, 134, 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1.

Chai, C. P. (2019), "Text Mining in Survey Data," *Survey Practice*, 12, 1–14. https://doi.org/10.29115/SP-2018-0035.

Chen, T., and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, T., He, T., Benesty, M., Kohtilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2020), *xgboost - Extreme Gradient Boosting*, R, .

Cimentada, J. (2019), *Download Data from the European Social Survey on the Fly R package version. 1.0.3.*, R, .

Cover, T., and Hart, P. (1967), "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Cunningham, D., Flicker, L., Murpy, J., Aldworth, J., Myers, S., and Kennet, J. (2005), "Incidence and Impact of Controlled Access Situations on Nonresponse," in *JSM Proceedings*, Proceedings of the Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Cunningham, P., Martin, D., and Brick, M. J. (2003), "An Experiment in Call Scheduling," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 59–66.

Dennis, J. M., Saulsberry, C., Battaglia, M. P., and Rodén, A.-S. (1999), "Analysis of Call Patterns in a Large Random-Digit-Dialling Survey: The National Immunization Survey."

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014), *Internet, phone, mail, and mixed-mode surveys: the tailored design method*, Hoboken: Wiley.

Douhou, S., Butt, S., Koch, A., and Briceno-Rosas, R. (2018), "ESS Round 9 Guidelines on Fieldwork Monitoring."

Durrant, G. B., D'Arrigo, J., and Steele, F. (2011), "Using paradata to predict best times of contact, conditioning on household and interviewer influences," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Journal of the Royal Statistical Society: Series A (Statistics in Society), 174.

Durrant, G. B., Groves, R. M., Staetsky, L., and Steele, F. (2010), “Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys,” *Public Opinion Quarterly*, 74, 1–36. https://doi.org/10.1093/poq/nfp098.

Durrant, G. B., and Steele, F. (2009), “Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys,” *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, [Wiley, Royal Statistical Society], 172, 361–381.

Eck, A. (2018), “Neural Networks for Survey Researchers,” *Survey Practice*, 11, 1–11. https://doi.org/10.29115/SP-2018-0002.

Esser, H. (1997), “Können Befragte Lügen?,” in *Soziologische Theorie und Empirie*, eds. J. Friedrichs, K. U. Mayer, and W. Schluchter, Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 261–283. https://doi.org/10.1007/978-3-322-80354-2_11.

European Social Survey (2018a), “ESS Round 9 Source Questionnaire,” ESS ERIC Headquarters c/o City, University of London.

European Social Survey (2018b), *ESS-9 2018 Documentation Report. Edition 1.2*, Bergen: European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC.

European Social Survey (2018c), *ESS-9 2018 Documentation Report. Edition 1.2*, Bergen: European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC.

Eurostat (2020), “Estimated average age of young people leaving the parental household by sex,” *Estimated average age of young people leaving the parental household by sex*, Available athttps://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-344495_QID_5A2654D3_UID_-3F171EB0&layout=SEX,L,X,0;GEO,L,Y,0;UNIT,L,Z,0;TIME,C,Z,1;INDICATORS,C,Z,2;&zSelection=DS-344495UNIT,AVG;DS-344495INDICATORS,OBS_FLAG;DS-344495TIME,2019;&rankName1=UNIT_1_2_-1_2&rankName2=INDICATORS_1_2_-1_2&rankName3=TIME_1_0_0_0&rankName4=SEX_1_2_0_0&rankName5=GEO_1_2_0_1&rStp=&cStp=&rDCh=&cDCh=&rDM=true&cDM=true&footnes=false&empty=false&wai=false&time_mode=ROLLING&time_most_recent=true&lang=EN&cfo=%23%23%23%2C%23%23%23.%23%23%23.

Fawcett, T. (2006), “An introduction to ROC analysis,” *Pattern Recognition Letters*, 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Felderer, B., Kirchner, A., and Kreuter, F. (2019), “The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias,” *Journal of Official Statistics*, 35, 93–115. https://doi.org/10.2478/jos-2019-0005.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (eds.) (2017), *Big data and social science: a practical guide to methods and tools*, Chapman & Hall/CRC statistics in the social and behavioral sciences series, Boca Raton, FL: CRC Press Taylor & Francis Group.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, 33. https://doi.org/10.18637/jss.v033.i01.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2020), *Lasso and Elastic-Net Regularized Generalized Linear Models*, R, .

Goyder, J. (1987), *The silent minority: nonrespondents on sample surveys*, Boulder, Colo: Westview Press.

Greenberg, B. B., and Stokes, L. S. (1990), “Developing an Optimal Call Scheduling Strategy for a Telephone Survey,” *Journal of Official Statistics*, 6, 421–435.

Groves, R. M. (2004), *Survey errors and survey costs*, Wiley series in survey methodology, Hoboken, N.J: Wiley.

Groves, R. M. (2006), “Nonresponse Rates and Nonresponse Bias in Household Surveys,” *Public Opinion Quarterly*, 70, 646–675. https://doi.org/10.1093/poq/nfl033.

Groves, R. M. (ed.) (2009), *Survey methodology*, Wiley series in survey methodology, Hoboken, N.J: Wiley.

Groves, R. M., and Couper, M. P. (1998), *Nonresponse in household interview surveys*, New York: Wiley.

Groves, R. M., and Heeringa, S. G. (2006), “Responsive design for household surveys: tools for actively controlling survey errors and costs,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Heck-Grossek, N., and Dardha, S. (2020), “Void of the Voiceless. An Analysis of Residents With a Language Barrier in Germany, France, and the United Kingdom,” in *The essential role of language in survey research*, RTI Press Publication No. BK-0023-2004: RTI Press, pp. 117–126.

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., Lyberg, L., and John Wiley & Sons (eds.) (2020), *Big data meets survey science: a collection of innovative methods*, Hoboken, NJ: John Wiley & Sons, Inc.

Hox, J., Blohm, M., and Koch, A. (2006), “The Influence of Interviewers’ Contact Behavior on the Contact and Cooperation Rate in Face-to-Face Household Surveys,” *International Journal of Public Opinion Research*, 19. https://doi.org/10.1093/ijpor/edh120.

Hsu, F. -h., Anantharaman, T. S., Campbell, M. S., and Nowatzyk, A. (1990), “Deep Thought,” in *Computers, Chess, and Cognition*, eds. T. A. Marsland and J. Schaeffer, New York, NY: Springer New York, pp. 55–78. https://doi.org/10.1007/978-1-4613-9080-0_5.

Hyafil, L., and Rivest, R. L. (1976), “Constructing optimal binary decision trees is NP-complete,” *Information Processing Letters*, 5, 15–17. https://doi.org/10.1016/0020-0190(76)90095-8.

Jäckle, A., Lynn, P., Sinibaldi, J., and Tipping, S. (2013), “The Effect of Interviewer Experience, Attitudes, Personality and Skills on Respondent Co-operation with Face-to-Face Surveys,” *Survey Research Methods*, 7, 1–15. https://doi.org/10.18148/srm/2013.v7i1.4736.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (eds.) (2013), *An introduction to statistical learning: with applications in R*, Springer texts in statistics, New York: Springer.

Juster, F. T., and Stafford, F. P. (eds.) (1985), *Time, goods, and well-being*, Ann Arbor, Mich: Survey Research Center, Institute for Social Research, University of Michigan.

Kalton, G. (1983), *Introduction to survey sampling*, Sage university papers series, Beverly Hills: Sage Publications.

Kern, C., Klausch, T., and Kreuter, F. (2019a), "Tree-based Machine Learning Methods for Survey Research," *Survey Research Methods*, European Survey Research Association, Vol 13, 73-93 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7395.

Kern, C., Weiss, B., and Kolb, J.-P. (2019b), "A Longitudinal Framework for Predicting Nonresponse in Panel Surveys," *Preprint*.

Kirchner, A., and Signorino, C. S. (2018), "Using Support Vector Machines for Survey Research," *Survey Practice*, 11, 1–14. https://doi.org/10.29115/SP-2018-0001.

Kolenikov, S., and Buskirk, T. D. (2015), "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification," Survey Methods: Insights from the Field (SMIF). https://doi.org/10.13094/SMIF-2015-00003.

Kopp, J., and Lois, D. (2014), *Sozialwissenschaftliche Datenanalyse*, Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-02300-3.

Krejčí, J. (2007), "Non-Response in Probability Sample Surveys in the Czech Republic," *Czech Sociological Review*, 43, 561–587.

Kreuter, F., and Müller, G. (2015), "A Note on Improving Process Efficiency in Panel Surveys with Paradata," *Field Methods*, 27, 55–65. https://doi.org/10.1177/1525822X14538205.

Kubat, M. (2017), *An introduction to machine learning*, New York, NY: Springer Science+Business Media.

Kuhn, M. (2014), "Futility Analysis in the Cross-Validation of Machine Learning Models," *arXiv*, stat.ML, 1405.6974v1.

Kuhn, M. (2019), *The caret Package*, R, .

Kulka, R. A., and Weeks, M. F. (1988), "Toward the development of optimal calling protocols for telephone surveys: a conditional probabilities approach," RTI International. P.O. Box 12194, Research Triangle Park, NC 27709-2194. Tel: 919-541-6000; e-mail: publications@rit.org; Web site: http://www.rti.org.

Laurie, H., Smith, R. A., and Scott, L. (1999), "Strategies for reducing nonresponse in a longitudinal panel survey," *Journal of Official Statistics*, Statistics Sweden, 15, 269–282.

de Leeuw, E., and de Heer, W. (2002), "Trends in household survey non-response. A longitudinal and international perspective," in *Survey nonresponse*, ed. R. M. Groves, New York: Wiley, pp. 41–54.

Lepkowski, J. M., Mosher, W. D., Davis, K. E., Groves, R. M., and Van Hoewyk, J. (2010), "The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey," *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, 1–36.

Lievesley, D. (1983), "Reducing Unit Non-response in Interview Surveys," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 295–299.

Lipps, O. (2008), "A Note on Interviewer Performance Measures in centralised CATI Surveys," *Survey Research Methods*, European Survey Research Association, Vol 2, 61-73 Pages. https://doi.org/10.18148/SRM/2008.V2I2.310.

Lipps, O. (2009), "Cooperation in Centralised CATI Household Panel Surveys - A Contact-based Multilevel Analysis to Examine Interviewer, Respondent, and Fieldwork Process Effects," *Journal of official statistics*, 25.

Lipps, O. (2012), "A Note on Improving Contact Times in Panel Surveys," *Field Methods*, SAGE Publications Inc, 24, 95–111. https://doi.org/10.1177/1525822X11417966.

Lipps, O. (2016), "Modelling Cooperation in an Address-register-based Telephone/Face-to-face Survey," *Field Methods*, SAGE Publications Inc, 28, 396–414. https://doi.org/10.1177/1525822X15617920.

Lipps, O., and Benson, G. (2005), "Cross-national contact strategies.," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 3905–3914.

Liu, M. (2020), "Using Machine Learning Models to Predict Attrition in a Survey Panel," in *Big Data Meets Survey Science*, eds. C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, and L. E. Lyberg, Wiley, pp. 415–433. https://doi.org/10.1002/9781118976357.ch14.

Luiten, A., Hox, J., and de Leeuw, E. (2020), "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys," *Journal of Official Statistics*, 36, 469–487. https://doi.org/10.2478/jos-2020-0025.

Luiten, A., and Schouten, B. (2013), "Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 176, 169–189. https://doi.org/10.2307/23355182.

Lynn, P., and Clarke, P. (2002), "Separating refusal bias and non-contact bias: evidence from UK national surveys," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51, 319–333. https://doi.org/10.1111/1467-9884.00321.

Malnar, B. (2021), *ESS International Bibliography 2003-2020. Based on Google Scholar Indexing. Includes 5,429 publications using ESS data.*, ESS ERIC WP11.

Massey, J. T. (1996), "Optimum calling patterns for random digit dialled telephone surveys," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 485–490.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019), "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning," *Bulletin of the American Meteorological Society*, 100, 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1.

McKay, S. (2019), “Can ‘Machine Learning’ Improve Our Understanding of Non-Response in ‘Understanding Society’?,” Essex.

Medway, D., Parker, C., and Roper, S. (2016), “Litter, gender and brand: The anticipation of incivilities and perceptions of crime prevalence,” *Journal of Environmental Psychology*, 45, 135–144. https://doi.org/10.1016/j.jenvp.2015.12.002.

Menard, S. (2002), *Applied Logistic Regression Analysis*, 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. https://doi.org/10.4135/9781412983433.

Müller, A. C., and Guido, S. (2016), *Introduction to machine learning with Python: a guide for data scientists*, Sebastopol, CA: O’Reilly Media, Inc.

Noack, M. (2015), *Methodische Probleme bei der Messung von Kriminalitätsfurcht und Viktimisierungserfahrungen*, Kriminalität und Gesellschaft, Wiesbaden: Springer VS.

O’Muircheartaigh, C., and Campanelli, P. (1999), “A multilevel exploration of the role of interviewers in survey non-response,” *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 162, 437–446.

Peress, M. (2010), “Correcting for Survey Nonresponse Using Variable Response Propensity,” *Journal of the American Statistical Association*, 105, 1418–1430. https://doi.org/10.1198/jasa.2010.ap09485.

Purdon, S., Campanelli, P., and Sturgis, P. (1999), “Interviewers’ Calling Strategies on Face-to-Face Interview Surveys,” *Journal of Official Statistics*, 15, 199–216.

Quinlan, J. R. (1986), “Induction of decision trees,” *Machine Learning*, 1, 81–106. https://doi.org/10.1007/BF00116251.

Robinson, J. P., and Godbey, G. (1997), “Time for Life: The Surprising Ways Americans Use Their Time.” https://doi.org/10.2307/2655174.

Rose, A. M. (1959), “Attitudinal Correlates of Social Participation,” *Social Forces*, 37, 202–206. https://doi.org/10.2307/2572962.

Sangodiah, A., Muniandy, M., and Heng, L. E. (2015), “Question Classification Using Statistical Approach. A Complete Review,” *Journal of Theoretical and Applied Information Technology*, 71, 386–395.

Schnell, R. (1990), “Computersimulationen und Theoriebildung in den Sozialwissenschaften.,” *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 42, 109–128.

Schnell, R. (1997), *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*, VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-97380-1.

Schnell, R. (2012), *Survey-Interviews: Methoden standardisierter Befragungen*, Studienskripten zur Soziologie, Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-19901-6.

Signorino, C. S., and Kirchner, A. (2018), “Using LASSO to Model Interactions and Nonlinearities in Survey Data,” *Survey Practice*, 11, 1–10. https://doi.org/10.29115/SP-2018-0005.

Singer, E., Frankel, M. R., and Glassman, M. B. (1983), “The Effect of Interviewer Characteristics and Expectations on Response,” *Public Opinion Quarterly*, 47, 68–83. https://doi.org/10.1086/268767.

Smith, T. W. (1983), “The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey,” *Public Opinion Quarterly*, 47, 386. https://doi.org/10.1086/268797.

Stanley, M., Roycroft, J., Amaya, A., Dever, J. A., and Srivastav, A. (2020), “The Effectiveness of Incentives on Completion Rates, Data Quality, and Nonresponse Bias in a Probability-based Internet Panel Survey,” *Field Methods*, 32, 159–179. https://doi.org/10.1177/1525822X20901802.

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., and Allred, S. B. (2019), “The End of the (Research) World As We Know It? Understanding and Coping With Declining Response Rates to Mail Surveys,” *Society & Natural Resources*, 32, 1139–1154. https://doi.org/10.1080/08941920.2019.1587127.

Stoop, I. (2005), *The Hunt for the Last Respondent*, Aksant Academic Publishers.

Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010), *Improving Survey Response: Lessons learned from the European Social Survey*, Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470688335.

Stoop, I., Koch, A., Loosveldt, G., and Kappelhof, J. (2018), “Field Procedures in the European Social Survey Round 9. Guidelines for Enhancing Response Rates and Minimising Nonresponse Bias.”

Sturgis, P., Williams, J., Brunton-Smith, I., and Moore, J. (2017), “Fieldwork Effort, Response Rate, and the Distribution of Survey Outcomes,” *Public Opinion Quarterly*, 81, 523–542. https://doi.org/10.1093/poq/nfw055.

Suthaharan, S. (2016), “Support Vector Machine,” in *Machine Learning Models and Algorithms for Big Data Classification*, Integrated Series in Information Systems, Boston, MA: Springer US, pp. 207–235. https://doi.org/10.1007/978-1-4899-7641-3_9.

Trappmann, M., Gramlich, T., and Mosthaf, A. (2015), “The effect of events between waves on panel attrition,” *Survey Research Methods*, 9, 31–43. https://doi.org/10.18148/srm/2015.v9i1.5849.

Trochim, W. M. K. (2005), *Research methods: the concise knowledge base*, Cincinnati, Ohio: Atomic Dog Pub.

Vandenplas, C., Loosveldt, G., and Beullens, K. (2017), “Fieldwork Monitoring for the European Social Survey: An illustration with Belgium and the Czech Republic in Round 7,” *Journal of Official Statistics*, 33, 659–686. https://doi.org/10.1515/jos-2017-0031.

Vicente, P. (2017), “Exploring fieldwork effects in a mobile CATI survey.,” *International Journal of Market Research*, blh, 59, 57–76.

Wagner, J. (2013), “Adaptive Contact Strategies in Telephone and Face-to-Face Surveys.,” *Survey Research Methods*, 7, 45–55.

Wang, K., Murphy, J., Baxter, R., and Aldworth, J. (2005), “Are Two Feet in the Door Better than One? Using Process Data to Examine Interviewer Effort and Nonresponse Bias.”

Weeks, M. F., Jones, B., Folsom, R., and Benrud, C. (1980), “Optimal times to contact sample households,” RTI International. P.O. Box 12194, Research Triangle Park, NC 27709-2194. Tel: 919-541-6000; e-mail: publications@rit.org; Web site: http://www.rti.org.

Weeks, M. F., Kulka, R. A., and Pierson, S. A. (1987), “Optimal Call Scheduling for a Telephone Survey,” *The Public Opinion Quarterly*, [Oxford University Press, American Association for Public Opinion Research], 51, 540–549.

Weidman, P. Z. (2010), “Do characteristics of RDD survey respondents differ according to difficulty of obtaining response?” *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 3956–3965.

Weisberg, H. F. (2005), *The total survey error approach: a guide to the new science of survey research*, Chicago: University of Chicago Press.

West, B. T., and Blom, A. G. (2017), “Explaining Interviewer Effects: A Research Synthesis,” *Journal of Survey Statistics and Methodology*, 5, 175–211. https://doi.org/10.1093/jssam/smw024.

West, B. T., and Sinibaldi, J. (2013), “The Quality of Paradata: A Literature Review,” in *Improving Surveys with Paradata*, ed. F. Kreuter, John Wiley & Sons, Ltd, pp. 339–359. https://doi.org/10.1002/9781118596869.ch14.

Würbach, A., and Zinn, S. (2019), “Using Paradata for Longitudinal Prediction of Participation Status. An Example of the NEPS Newborn Cohort.,” Zagreb.

XGBoost Developers (2020), “Final Words on XGBoost,” *XGBoost Documentation*, Documentation, , Available athttps://xgboost.readthedocs.io/en/latest/tutorials/model.html?highlight=push%20extreme.

Zhang, D., and Lee, W. S. (2003), “Question classification using support vector machines,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’03*, Toronto, Canada: ACM Press, p. 26. https://doi.org/10.1145/860435.860443.

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., and Goyal, H. (2018), “Opening the black box of neural networks: methods for interpreting neural network models in clinical applications,” *Annals of Translational Medicine*, 6, 216–216. https://doi.org/10.21037/atm.2018.05.32.