7

Data Subjects: Calibrating and Sieving

Baki Cakici and Evelyn Ruppert

Introduction[1]

The problem, however, is to get the respondent to answer these questions.[2]

Who are the subjects of data practices? How do data practices configure the capacities of subjects to engage and participate in their categorisation and become part of a population? These are questions this chapter turns to by first assuming that subjects do not pre-exist data practices but come into being through them (Ruppert, 2011). The data practices analysed in the foregoing chapters, such as those that make up administrative registers and surveys, involve different relations to what this chapter refers to as data subjects. Whether implicit or explicit, data practices that encode people into categories, for example, interact and engage with subjects in different ways. And, in doing so, data subjects come into being through varying relations, interactions, and dynamics between human and technological actors that make up data practices. This is distinct from usual understandings of data subjects, who are typically conceived as people who have a passive entitlement to their personal data and privacy, a right that is regulated by the state (Guild, 2019: 268). Similarly, it is different from an understanding

that conceives of data subjects as 'data doubles' (Haggerty and Ericson, 2000), which implies that data are simply digital duplicates rather than the products of subjects' relations with digital technologies.[3] Rather, this chapter explores how data subjects neither pre-exist nor are passive but shaped through data practices that configure their capacities to intervene, challenge, and influence how they are then categorised and become part of a population. Such configurations and capacities are variable and contingent because of different sociotechnical relations and data practices that make up methods; that is, data subjects are multiple, a point we demonstrate below through the analysis of two distinct data practices: calibrating and sieving.

A key aspect of the configuration of capacities that bring different data subjects into being concerns how data practices are organised and influenced by problematisations. For instance, as expressed in the opening quote, getting subjects to answer is a problem that is said to be evident in a general decline in response rates to paper questionnaires. This decline is usually attributed to people being overburdened by numerous state data collection activities or their concerns about privacy and confidentiality. Another cited cause explored in Chapter 4 is that certain groups, such as refugees and homeless people are identified as difficult to locate, contact, interview, and persuade to participate in data collection methods and thus 'hard-to-count'. However, even when subjects answer questionnaires, their responses can be a further source of critique. While expected to reveal themselves truthfully, subjects are also understood, in some cases, to answer strategically and subversively, for example, by claiming unrecognised or unauthoritative categories.[4] Many efforts are thus directed at improving the reliability of

responses, which often involve a tension between opening and closing the possibilities of how a subject can respond to a question:

> The value of open-ended questions is that they offer the respondent the right of total self-expression. The disadvantage is that the subsequent coding of responses and their allocation into a meaningful classification for output becomes more difficult and costly.[5]

One such example captured media attention in the UK in the wake of the 2001 census of England and Wales when more than 390,000 respondents declared 'Jedi' as their religion in response to a newly introduced optional question on religious beliefs. While the UK Office for National Statistics (ONS) categorised what was considered a subversive response under the 'no religion' category, the response was referenced in subsequent parliamentary debates on the future of population censuses and the inaccuracy of questionnaire-based methods. These are just a few of the problematisations of subjects whose self-elicited answers can also be influenced by how questions are worded or whether questions are self-completed or involve an enumerator.[6]

Such problematisations of data subjects come to inform and configure data practices that make up method experiments that engage with digital technologies as possible solutions. While also driven by problematisations of data quality, cost, and timeliness, it is how method experiments are offered as solutions to the (non)responsiveness and truthfulness of subjects that this chapter considers. That is, such problematisations of subjects' very capacity to act and influence (or subvert) how they are categorised, we argue, inform the development of data practices that are offered as solutions. We interpret two

such solutions – data practices that calibrate and sieve – and argue that they constitute different 'forces of subjectivation' (Cakici and Ruppert, 2019).

In brief, our conception builds on Foucault's (1982) formulation that subjects are capable of reflection, self-formation, and are engaged in struggles against direct domination that involves a tension between governing and technologies of the self. It is a power relationship that requires that a person is capable of acting and who, when faced with a relationship of power, engages with 'a whole field of responses, reactions, results, and possible inventions' (Foucault, 1982: 789). In this way, Foucault connected subjection and subjectivation to capture that power is not possessed but is a relation and process (Cremonesi et al., 2016). This relation and tension between governing and technologies of the self are well captured in Foucault's conception of subjectivation:

> On the other hand, a power relationship can only be articulated on the basis of two elements which are each indispensable if it is really to be a power relationship: that 'the other' (the one over whom power is exercised) be thoroughly recognized and maintained to the very end as a person who acts; and that, faced with a relationship of power, a whole field of responses, reactions, results, and possible inventions may open up (Foucault, 1982: 789).

The tension within a relationship of power is captured in a distinction suggested by Balibar (1991) between being a subject *to* power and a subject *of* power. Being a subject to power means to be dominated by and obedient to a sovereign. However, when a subject submits to power this opens the possibility to be subversive and be a subject of power. Regarding the latter possibility, this is what distinguishes being a citizen: one who is both a subject to and subject

of power, where obedience, submission, and subversion are always-present potentialities (Isin and Ruppert, 2015).

It is in this sense that the data practices that make up method experiments can be conceived of as forces of subjectivation. They are forces of power not in the sense that they determine but rather, through the different sociotechnical relations that make them up, differently configure the capacities of subjects to act in how they are categorised and become part of a population. For the data practices that make up methods require the actions of subjects – whether through the selection of a tick box or the entry of a location in a free-text field or the writing of a tweet – who participate in their subjectivation and categorisation. They can act in obedient and submissive ways and simply respond as expected and required or they can invent, subvert, and resist their subjectivation and perform as citizens including not participating or submitting to the data demands of governing authorities (Isin and Ruppert, 2015). As such, changes in data practices reconfigure the possibilities and potentials of acting and performing as citizens.

It is regarding this potential that method experiments can also be inventive of new forms of acting when they come into play and can also change initial problem formulations. For when put into action, the interactions and dynamics between human and technological actors are not determining but contingent. As Neyland and Milyaeva (2016) note in relation to market interventions, problems are not settled and given but often reworked, transformed, or lead to further problems. From climate change to vaccines, problems, solutions, and interventions are entangled and dynamically reformulated.

This conception of forces of subjectivation is taken up in this chapter to analyse two method experiments. They are considered as experiments insofar as they involve pilot

projects and the testing of innovations in methods that need to be proven not through argumentation but demonstration (Ruppert and Scheel, 2019). The analyses interpret the data practices that make up method experiments as sociotechnical and contingent in relation to how they configure, enable, or constrain how subjects might act. 'Calibrating responses' examines some of the data practices involved in digital censuses and how they seek to maximise and guide the responses of subjects. 'Sieving tweets' focuses on data practices involved in experiments with Twitter for generating 'live' data about the dynamics of student internal migration. In both cases, we examine how classifying and encoding subjects, as defined by Desrosières (1998), involve different forces of subjectivation that seek to maximise the obedience and submission of subjects. The conclusion reflects on these forces to consider the consequences of data practices for the possibilities of subjects to act as 'data citizens' (Ruppert, 2018) in how they are categorised and encoded as part of a population.

Calibrating Responses

In 2011, following years of design and development, Estonia tested and conducted its first e-census. Reporting on the outcomes, Estonian statisticians declared that the country 'reached international premiere league' in that 'all people could fill out their personal questionnaire online' with the result that the country 'set the world record' with 66 per cent of respondents using the e-census (Tiit, 2013, 2015). This evaluation of success reflects the relation of the e-census to similar NSI method experiments with digital, online, or e-censuses (generally referred to as digital censuses) over the past decade. As one solution to the problems of paper questionnaires, these experiments are at various stages of design and implementation and

circulate in reports, international presentations, and demonstrations within and beyond EU NSIs. Rather than inventions of individual NSIs, problems and solutions are being identified, developed, repeated, referenced, debated, and contested and travel and circulate in and through the transnational field of statistics (Scheel, Grommé, and Ruppert, 2016). As such, the field includes states that make up the EU as well as those that form part of the UNECE. The examples analysed here are understood to be part of this field and through which national statisticians introduce, demonstrate, and defend new data practices as well as compete to set 'world records'.

Returning to the report on the Estonian e-census, statisticians noted that achieving a high online response rate involved an 'information and motivation campaign' that explained how a tachometer would track the volume of active respondents completing the census. One report described how the use of online enumeration rose to unexpected levels, despite the tachometer warning that the platform was experiencing a high volume of activity and that respondents might best do their submission later. Because of high volumes, the time required for responding was doubled, which further exacerbated online congestion. Customer support was subsequently unable to answer all incoming questions and internet services were interrupted at one point for about half an hour. Measures were taken to improve the situation on the following day and no further major technical setbacks were experienced. After this intense start-up, when approximately 50,000 people completed the online questionnaire in one day, levels dropped to 20,000 over the final two weeks (Statistics Estonia, 2012).

This account highlights some valuations and considerations related to NSI method experiments with digital censuses, which are more generally positioned as part of a broader move to 'digital government'. For example, Estonian statisticians

described the e-census as 'essentially, a grand IT project' (Statistics Estonia, 2012) that is part of what the government refers to as e-Estonia:

> Estonian people are used to thinking that Estonia is an e-country. We have an e-state and a wide range of e-services. Sometimes we worry whether other countries are overtaking us in the e-race. It is, of course, difficult to measure a country's e-capability, as there are no uniform indicators in this area. However, the census reinforced the notion of e-Estonia, which is positive. Not only because we are proud to be e-Estonia, but also because the active participation in the e-census will probably help us to conduct the next census with lower costs and greater efficiency (Oopkaup and Servinski 2013, 17).

Reflecting on the case of e-Estonia, a UK report described this as transforming government through technology and 'the relationship between citizens and the State – putting more power in the hands of citizens and being more responsive to their needs' (UK, 2017: 21). While oriented to numerous objectives, such as lower cost and efficiency, accounts of digital government, and more specifically of the Estonian e-census, proclaim the possibilities of digital technologies to establish a new relation between subjects and the state. However, the data practices that make up digital censuses configure this relation in particular ways that enable, constrain, and configure the forces of subjectivation and how subjects are categorised and become part of a population. Rather than simply tools, technologies such as the live tracking of responses and tachometers are part of an array of sociotechnical actors that make up these forces.

Such an array of forces is exemplified in Australia's design of a digital census. According to a statistician in a presentation made at a UK international conference in 2014, rather than an

online census, a digital census does not simply use digital technologies such as the internet to collect data and disseminate results.[7] It means to do all aspects of the census digitally. Their presentation reflected on the Australian Bureau of Statistics' (ABS) plans for its first 'digital census' in 2016, which they said would involve a 'transformation' rather than simply 'translation' of a paper questionnaire into digital format. It would involve a move from digital publishing to digital transacting and interacting with subjects at all stages of enumeration, and a responsive approach that would make data collection adjustments in near 'real-time' based on field intelligence and response rates (Australian Bureau of Statistics, 2015). A central management centre would achieve this by digitally monitoring a range of management information, including online response rates, paper form requests and returns, and social media. For example, when the response rate of an area lagged others, then a variation to the enumeration approach would be designed, reviewed, and actioned.

The statistician's presentation conveyed how relations between a digital census and subjects are understood. They are relations that can be interpreted as involving entangled human and technological relations that emerge through a dynamic call-and-response between subjects and technologies. While no method can direct subjects to one and only one way of acting, the data practices that make up the digital census are arranged to manage and guide how subjects act. In other words, they anticipate how a subject might act and identify, and seek to manage, direct, and channel those possibilities. It is in this way that a digital census anticipates subjects. As other researchers have elaborated, anticipatory logics underpin both governing and technical practices and are speculative regimes and forces (Adams, Murphy, and Clarke, 2009; Ratner, 2019). Anticipatory and pre-emptive logics, for example, have

been explored in relation to security and surveillance (Aradau and Blanke 2018). However, these studies address anticipatory logics involved in the analysis of data rather than the practices that configure relations to subjects. As developed below, the data practices that make up digital censuses anticipate how subjects might act and do so dynamically through what we describe as calibration.

For example, the ABS statistician, in their presentation to the UK international conference, described how putting a questionnaire online does not merely change the relation to subjects but transforms it into an interaction that is 'easy, responsive, fun.' The proposed design would do this by providing more information through pop-up windows to guide correct responses; drag and drop techniques to facilitate the ease of completing questions; assistance prompts to guide experience such as supplementary questions; and images and summary compilations that visualise responses so that they can be verified by subjects. The Estonia e-census also included help texts and 'soft and strict logical controls' to 'prevent or highlight the majority of logically impossible responses' (Statistics Estonia 2012, 3).

For the Australian digital census, the management of relations also extended to a 'field force' of workers who would use digital technologies to better capture and monitor subjects. By digitally monitoring progress through handheld devices, constant feedback on operational progress and instructions would be fed back to workers to optimise their activity and highlight problem areas in response rates. Social media platforms such as Twitter would also be used by workers to communicate experiences to each other so that problematic subjects and areas could be better targeted. Similarly, Estonia's e-census included 'The Survey Fieldwork Information System (VVIS)', which created work lists for enumeration areas, managed the roles of census team members, and monitored interviews amongst other things (Statistics Estonia 2012, 3).

All of these features were implemented in Australia's 2016 digital census and Estonia's 2011 e-census. Through numerous data practices, subjectivation was transformed into an interactive and live process of calibrating the responses of subjects by prompting and guiding them and making the process fun and easy and thereby maximise their submission to the census. Subjects who did not submit or obey in ways anticipated, were then targeted either by digital techniques such as prompts or by enumerators deployed through offline modes in the field. Significantly, in contrast to paper questionnaires, this was conceived of as happening in 'real time', rather than through long processes of testing, piloting, and field worker feedback. With digital censuses then, relations between digital technologies, central management, and field workers that make up the method are organised by data practices that are dynamic, recursive, and responsive.

At the same time, the humans and technologies that participate in digital censuses extend to multiple other data practices such as those comprising administrative registers, self-completed paper questionnaires, and interviews conducted by enumerators using digital questionnaires on laptops. For example, in Estonia, registers were used in various ways such as to pre-fill some answers on questionnaires and supplement results when data was missing (Statistics Estonia, 2012).[8] In these ways, digital censuses are part of broader method assemblages that consist of data practices involving numerous technologies, rules, things, concepts, and people.

Producing New Problematic Subjects

At the 2015 annual meeting of the UNECE Group of Experts on Population and Housing Censuses, a statistician from the UK ONS noted that his office had learned much from international colleagues and their census practices. He noted that

international practices had influenced the UK's decision to introduce a major change in what he referred to as the '2021 Census Transformation Programme': that censuses would be conducted 'online first' and supplemented by multimode follow-up methods to capture non-responding households.[9] The statistician noted that the online census would also go beyond the simple translation of a paper questionnaire to incorporate many of the elements adopted in the Australian digital census such as contextual assistance for subjects to complete questions; detailed drop-down boxes to reduce coding; comprehensive validation within and between questions; and the design of questions to fit smaller screens so that subjects could respond using handheld devices (ONS, 2015a).

Over time, this initial conception of the ONS Census Programme lead to the design of an online census that was promoted as a 'digital-first approach' and which would be 'easy to complete, and rewarding for respondents, so 70% provide data without follow-up' such that '75% of responses [are] provided online, and assistance provided to those who need it, to make this the most inclusive census ever' (HM Government 2018, 3). It would adopt smart type-in options and 'search-as-you-type' capabilities and functions such as routing, validation, and guidance. Additionally, through multi-channel and multi-lingual communications, community engagement, and the advice and help of field force and contact centre staff, the design would 'ensure people can tell us how they wish to identify themselves' (10). These and other sociotechnical arrangements would make up the many 'interactions with the census respondent'.

The validation and smart type-in features of digital censuses referred to above are made possible by the generation of paradata, which is a type of metadata.[10] Rather than being descriptive of the practices through which data has been

generated as in traditional metadata, paradata constitutes 'process' data on a subject's digital actions.[11] It is sometimes referred to as big data because it is generated in 'real-time,' and in large volumes that require processing by algorithms. It includes data on devices being used; timestamps; which buttons (help, back, forward) are being clicked and when; changes subjects make to answers; and so on (Statistics Austria, 2015). For each, inferences can be made about myriad issues such as individual subjects and groups who do not submit to the census in ways anticipated and desired because of one of these design elements. In these ways, paradata involves tracking the relation between the digital census and the subject through metrics about data collection and are part of a 'data driven approach,' which informs strategies for increasing response rates and the submission of subjects. It is a by-product of digital technologies that can be put in the service of better calibrating responses.

Using 'smart' technologies such as autocomplete, the data practices of digital censuses thus operate like commercial digital platforms. Indeed, one justification for digital censuses is that subjects regularly engage with digital platforms for both public and commercial purposes and thus have the familiarity and skills necessary. At the same time, digital censuses adopt many of the elements of the user interfaces that make up these other platforms – especially those of Google, Facebook and Amazon – and which are criticised for channelling choices and directing queries (König and Rasch, 2014; van Dijck, Poell, and De Waal, 2018). While user interfaces such as Google's query function appear neutral, autocomplete suggestions anticipate and predict what users want to know and direct queries through suggestions. Like smart type-in, logical controls on entries, and assistance prompts, autocomplete is intended to make searching faster and easier and produce optimal results. In these ways, digital censuses

incorporate practices innovated and designed by private technology companies. As such they also adopt similar logics, especially those advanced by data science, which seek to tame, control, and guide the actions of subjects through a new science of societies that challenges existing forms of data and knowledge such as that generated by traditional methods and practices of national statisticians (Grommé, Ruppert, and Cakici, 2018).

While all data practices variously channel and direct answers of subjects through techniques such as tick boxes on questionnaires, digital technologies do this in ways that are less evident and work in the background to increase submission by reducing the possibilities of intervening and subverting. Like internet platforms that espouse process data as working in the service of a better and faster customer service, so too is paradata mobilised in the service of better and faster responses to digital censuses. Through both the identification and subsequent capture of evasive, hard-to-count subjects, calibrating aims to normalise them through techniques that entice responses through fun elements and gamification and that discipline by anticipating and preventing illogical or unrecognised responses. In this way forces of subjectivation configure capacities and possibilities for acting.

However, while an online census was promoted by ONS for its capacity to ensure correct responses from subjects, it also produced new problematic subjects. Four groups of problematic subjects were anticipated based on their expected access to and/or willingness to use the internet to digitally engage with government via the internet (Figure 7.1). Problematic subjects – like hard-to-count subjects discussed in Chapter 4 – were differentiated according to several criteria. For each group, their related sociodemographic characteristics were identified (age, location, etc.) as well as reasons for

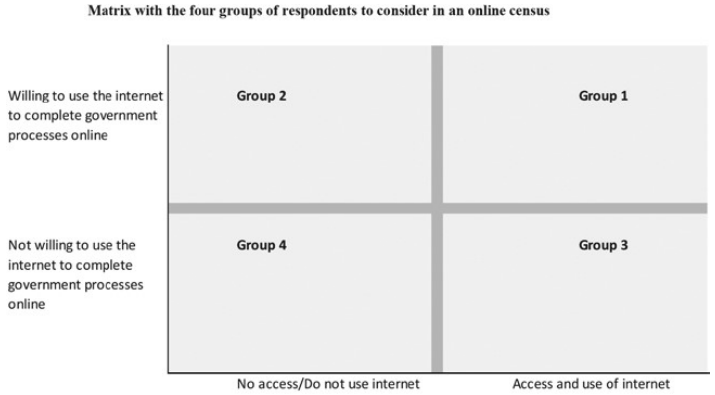

Figure 7.1 Categorisation of Respondents[a]
[a]Source: ONS, 2015a

being unwilling to digitally engage (lack of trust, internet security, etc.). In this conception, a digital divide was conceived not simply between who does or does not have access to the internet, but as divisions that occur along various combinations of identification such as where someone lives and their age. These characteristics were used to calculate the numbers of likely hard-to-count subjects and their relative concentration in different geographic areas. Response rates and patterns could then be tracked in these areas and direct follow-up field activities organised when targets were not being met so as to increase the number of subjects who submit to the census.

Such management involved offline modes as demonstrated in the ONS's test of its online census in 2017. The test was designed to evaluate options for maximising responses, self-completion, and the quality of responses. One element evaluated was the introduction of an 'Assisted Digital Service' to reach the 'more than 10% of UK adults who have never used the internet' and recognition that '21% of the population

lack basic online skills' (Bexley, 2017). The service involved setting up computer terminals in a local library with librarians to assist subjects in completing an online questionnaire. The decision on the design of the 2021 census included this service, which involved 'trusted suppliers who have the staff, premises and technology' to help respondents as well as the organisation of 'completion events' to stimulate response rates (HM Government 2018, 5).

Subjectivation thus involves data practices that anticipate how subjects might act and then calibrate how they do act through the ongoing process of digital management and directing. That is, a digital census does not simply involve deploying digital technologies but managing their operation and the performance of subjects in relation to them as live processes of subjectivation. However, as illustrated above, the design of a digital census is generative of new problematic subjects and calls forth management solutions in the form of new actors (librarians, enumerators), sites (libraries and computer terminals), and data (paradata), which all participate in subjectivation. All of these participate in the forces of subjectivation and inventive of data subjects who do not pre-exist but come into being through data practices that configure the relations, interactions, and dynamics between human and technological actors.

Yet, management is not only necessary to direct subjects, but also to address the instability and vulnerabilities of digital technologies. While this can take many forms, such as a change in an operating system as noted in the next section, a dramatic example was the disruption to the Australian digital census website, which suffered a mass outage and was shut down for 43 hours during the 2016 enumeration (MacGibbon, 2016). Attributed to a Distributed Denial of Service Attack (DDoS), the failure led to a major inquiry into cybersecurity and the

close partnership between ABS and IBM.[12] Loss of public trust and confidence were widely noted as a major consequence but what the incident points to are the contingencies of digital technologies. Not only are they subject to operational failures, but other forms of subversion because of the introduction of new technological and human actors that reconfigure those possibilities. Additionally, such contingencies reduce the submission of subjects to the digital census and, in turn, desired response rates.

In response to a recommendation in that report, ABS established an Independent Assurance Panel (the Panel) to secure trust in census operations and the quality of data generated. Rather than an assessment of individual features of the digital census, their assessment was that the 2016 census produced data of comparable quality to previous censuses and 'is useful and useable' (Census Independent Assurance Panel to the Australian Statistician 2017, iii). The relevance of the digital mattered only in relation to the DDoS rather than all the other proclaimed benefits and operational features detailed above. While internal reviews may well focus on this, the public response concerned the security of the digital census and confidence in its quality, and the degree to which subjects submit to and act in ways anticipated. As we explore in the next section on method experiments with Twitter data, the dynamics of sociotechnical relations, and the contingencies of data practices that configure subjectivation, can lead to other unexpected outcomes.

Sieving tweets

In this section, we explore the dynamics of subjectivation in relation to one experiment, an ONS pilot project that sought to use Twitter data to investigate how populations move within

the UK (ONS, 2015b). In 2014 and 2015, ONS statisticians experimented with a method to identify patterns in when and where users create Twitter posts based on aggregated data collected from publicly available Twitter profiles.[13] Their driving assumption was that if tweets originated from different places at different times throughout the year, it would be possible to identify a pattern, and infer underlying reasons for why people move from one place to another. They argued that this would be an improvement over subjects declaring their mobility patterns on questionnaires as it would avoid false reporting and underreporting (i.e., where respondents either provide a wrong address, or provide only one address when they occupy several). The statisticians believed that it could also provide more timely statistics about how people move between addresses throughout the year.

This section explores how this method experiment involved sieving as a data practice and force of subjectivation. Like the previous example, the experiment was offered as a potential solution to problematised subjects, in this instance that of higher education students. They are deemed hard-to-count because of their irregular movements between universities and multiple residences within the academic year, which makes it difficult to encode them to a usual residence (on the problematisation of mobile people as 'hard-to-count' see Chapter 3). As elaborated below, sieving involves repurposing tweets to filter and sort subjects and then infer and enact the category of migrating students. In distinction to calibrating, which iteratively incites, disciplines, and interacts with subjects to participate in their categorisation, sieving is a force of subjectivation that does not engage with subjects but categorises them based on repurposing big data about their conduct. That is, rather than guiding subjects, sieving eliminates the possibilities of subjects to act in – or even know – how they were categorised and the possibilities of their intervention.

The experiment more generally held the promise of providing more timely statistics that reflect lived experiences and which do not rely on elicited (and unreliable) responses from subjects. By repurposing the data traces of Twitter users, the pilot followed method experiments both within NSIs and academic research that engage with social media platforms such as Facebook profiles and Twitter posts to infer statistics on geography, language, and sometimes even gender and ethnicity (Liu and Ruths, 2013; Mislove et al., 2011; Mocanu et al., 2013; Nguyen et al., 2013; Sloan et al., 2015). These method experiments, which involve digital technologies, big data, and new analytics, diverge most significantly from paper questionnaires in that subjects do not self-identify. Rather than data from 'registers of talk' such as those of traditional methods, these experiments use data generated by platforms that are 'registers of action' (Marres, 2017). Subjects' identifications are inferred from data traces of their actions and collected for other purposes and constitute a different form of subjectivation. For one, subjects can neither opt-out or subvert inferences, but, as we detail below, through various adjustments to how they interact with platforms, they can engender new problematisations.

The method experiment involved several stages beginning with investigation of the free-text location field included in Twitter profiles. After a brief study, the statisticians in charge concluded that the text field is an unreliable data source as users seemed to use it in different ways, sometimes leaving it blank, and sometimes subverting the intended use by filling it with fictional places. The free-text field provided the potential for subjects to act in ways that subverted and were not compatible with the strict geographical definition of location necessary for the pilot project. As an alternative, the statisticians decided to concentrate solely on tweets that include GPS coordinates as these messages, also known as geolocated tweets, provide standardised data about the location from

which a tweet was posted. These were much easier to analyse using existing statistical methods, and less prone to the kinds of uncertainties introduced by users. However, they made up a fraction of the total number of tweets, and many were posted by the same users. Furthermore, GPS coordinates were linked to a much broader sociotechnical arrangement consisting of satellites, sensors, and mobile devices and generated a new set of unanticipated issues and different problematisations of subjects as we outline below.

To eliminate tweets that did not include GPS coordinates and thereby focus on a desired subset, the statisticians engaged in the data practice of sieving. Kockelman's (2013) conceptualisation of sieving in algorithmic devices shows how sieves have desires built into them; they retain a set of 'desirable' elements while allowing the 'undesirable' to disperse. This process was evident in the separation of tweets depending on the availability of the GPS coordinates, where the geolocated tweets – constituting a smaller volume – were gathered for further analysis and the rest were discarded. Such procedures were repeated with different sieves, for example one that allowed the removal of Twitter bots (accounts that post exceptionally high numbers of tweets in relation to the rest of users). Another sieve was necessary when the statisticians discovered that two sets of data they used, one purchased from a data reseller and another obtained using the Twitter API,[14] included duplicates because there was an overlap in the dates when the data were collected. While the work of sieving involved separating tweets in both cases, its significance was that it transformed undifferentiated collections into a potential source of data for inferring categories of subjects using existing statistical methods. In so doing, rather than engaging the desires of subjects in categorisation, sieving materialised categories that reflected the preferences and desires of statisticians for reliable and verifiable geolocations.

Although tweets in a chosen subset could now be linked to a geographical location using GPS coordinates, the stream of tweets for each user still needed to be translated into 'significant locations', namely work and home. To perform the translation, the statisticians used a clustering algorithm called DBScan, which arranged the stream of tweets for each user into clusters of nearby data points. Next, they used a set of rules about the time of day and frequency of posts to infer whether the assigned locations could be considered the home or the workplace of the posting user (see ONS (2012) for a detailed description of the method). Finally, they compared the positions of the tweet clusters to the borders of local authorities, and they flagged those that appeared in different local authorities from one month to the next as instances of internal migration. Using this analysis, the statisticians quickly detected a 'strong signal' coinciding with the cycle of the academic year. The signal indicated that in local authorities with high proportions of students, the volume of tweets seemed to decrease in June and increase again in September and October. Based on this finding, they concluded that the data could be used as an indicator of student mobility, movements that were not possible to detect using any existing data sources.

The production of dominant tweet clusters is another example of sieving in action. The algorithm (DBScan) converts a larger set of tweets into a much smaller one by allowing closely located tweets to pass and be included while blocking and discarding more dispersed ones. Which tweets are allowed to pass or are discarded are determined individually for each Twitter user, that is, a different sieve is used for each user, but the tweets themselves, and the location data they contain, remain unchanged throughout the process. In other words, the algorithm performs as a sieve by neither changing which tweets it catches, nor which ones it lets through.

While the data practice of sieving tweets led to inferring and in turn enacting the category of migrating students by repurposing existing Twitter user data, it also eliminated the possibilities of subjects to act in – or even know – how they were categorised and the possibilities of their intervention. To demonstrate the effect in action, we can consider the final inference that enacted the student migrant population. As noted previously, higher education students are often problematised subjects because their movements between universities and other residences within the academic year make it difficult to encode them in a usual residence (see discussion in Chapter 3). For example, statisticians have long argued that population counts conducted at different times in the same geographic area can display high variations if the size of the student population is sufficiently large (Duke-Williams, 2009; Mitchell et al., 2002). It is in the context of this problematisation that the statisticians on the pilot project came to recognise and identify a solution: by converting Twitter posts into geographic indicators the mobility of Twitter users could be inferred. That is, it was in relation to a well-known and debated problem that the pilot project invented a solution which could be legible and recognised as useful to produce statistics. It did so through the further stabilisation of the notion of student mobility, where studying involved living away from home while remaining connected to a home that exists in another location. The role of sieving as a data practice in this configuration is that it generated a potential solution to a problem by inferring and enacting the category of the migrating student.

Detecting and inferring student migration was a promising result for the pilot project as it solved the problem of categorising a hard-to-count mobile student population. However, the statistician in charge of the project noticed a significant decrease in the number of data points at a particular date in

the one-year sample of Twitter posts. After a period of investigation, they found out that the date of this decrease coincided with the release date of iOS 8 (an operating system used by Apple devices). Further investigation pointed to a change in the default settings in the operating system for location sharing, meaning that on that date many devices stopped reporting their locations, and thus disappeared from the dataset. This disappearance led the statistician to characterise the dataset as volatile, that is, unreliable and prone to sudden changes, and ultimately unsuitable as a data source for official statistics. In other words, problematic subjects were replaced by problematic, unreliable, and volatile technological actors.

While complications that arose when using GPS data for population data were easier to anticipate and handle for the statisticians, the GPS coordinates were thus also linked to a much larger method assemblage, a hinterland of actors consisting of networks of satellites, sensors, and mobile devices, all of which generated a new set of unanticipated issues. In this instance, the data practices were contingent due to their dependency on this assemblage, where changes in software release schedules or operating system settings of Twitter users, could jeopardise the otherwise stable results.

When the chief statistician described the data source as 'volatile', the description captured the contingencies of forces of subjectivation. In the pilot, using GPS coordinates to overcome the challenges of determining a location through free-text fields exposed other dependencies beyond the control of the project. At stake was the possibility of being able to anticipate technological actors; that is, even if the sharp decrease in user numbers could be tied to a single event this time, a similar change in the future might be impossible to anticipate, explain, or even to detect. Configuring subjectivation, in other words, was beyond the reach of their method as it was part of a widely

distributed assemblage of infrastructures and temporalities. In these ways, forces of subjectivation involve not only configuring, anticipating, and remediating the acts and actions of human subjects, but also those of technological actors.[15]

The Twitter pilot began as a method of a more 'live' tracking of mobility by sieving geolocated tweets to produce categories from clusters of data points made possible by a highly technical analysis. For us, it demonstrated how subjectivation is differently configured by data practices, but also that its force is the product of the interactions and dynamics between human and technological actors, including categories, software, algorithms, and digital devices. While the data practice chosen by statisticians inferred and enacted the category of migrating students, it arose from the complex interplay between location categories such as home and work, software settings, release schedules, and study design as well as the actions and inactions of subjects.

So, while the data practice of sieving was a solution to the problem of categorising migrating students, it was generative of a series of new problems. Subjects were problematised for their use of a free-text field, which generated unanticipated categories or interpretations. While GPS coordinates were identified as a solution, this made the method vulnerable to technical forces of operating systems involving actions beyond their control or knowledge. In these ways, while reconfigurations of forces of subjectivation may solve one set of problems, they can also be generative of new ones.

Conclusion

This chapter covered just a few examples of data practices that configure the capacities of subjects to engage and participate in their categorisation and how they become part of

a population. It highlighted that while cost, time, efficiency, and quality are key objectives of method experiments, they also are directed at reconfiguring how people are subjectified to meet desired ends through data practices that are not linear but recursive and dynamic. From the iterative calibrating of responses of digital censuses to the repetitive sieving of tweets, data practices work to minimise the subversive and maximise the submissive actions of subjects.

This objective was exemplified in problematisations of subversive or hard-to-count subjects such as those who answered Jedi in response to the 'no religion' question of the 2011 census of England and Wales. While a digital census was offered as a possible solution, by reconfiguring the forces of subjectivation, new hard-to-count subjects were anticipated and produced due to the introduction of digital technologies. In this regard, solutions are inventive of new possibilities for subjects to act, be excluded, or problematised. This is in part because data practices such as calibrating and sieving introduce new actors, such as the assumptions, objectives and biases of platforms and the decisions of operating system owners. However, rather than simply a question of reducing the potential of subjects to act, we have attended to how data practices differently configure their subjectivation, which can be anticipated and guided but not settled in advance.

Yet, there is another consequence. Methods not only configure the capacities of subjects to obey, submit, and subvert, they also configure their object, that is, the populations that are enacted. While populations have historically been understood as relatively stable objects that only require periodic measurement, the method experiments we have analysed enact them as fluid and modulating (Ruppert, 2012). In other words, new kinds of populations and modes of intervention are also

invented. Furthermore, while typically based on self-elicited social categories, some experiments infer identification categories and populations from the data traces of subjects generated by their actions in relation to digital platforms. In these ways, not only do methods produce their subjects and their agential capacities, but also the very object of population is transformed.

Data from digital platforms and mobile devices are also potentially transformative of the how European population statistics may be produced in the future. Method experiments such as those with Twitter – or mobile phones (e.g., see Ruppert and Scheel, 2019) – introduce big data that are transnational in their generation and ownership. Given that European population statistics are largely generated by and reliant upon national statistical institutes, big data introduce the prospect of transcending national borders to produce European level statistics. That is, rather than harmonising and assembling national data, European statistics could be based on transnational data. Since this data is owned by multinational corporations, European level governance and negotiation may be necessary to secure access if experiments are to lead to the production of internationally comparable population statistics.[16] Furthermore, if, as proposed in Chapter 1, statistics help to constitute what is the population and who are the people of Europe, then big data could be a key political technology through which the EU could possibly constitute its public and secure its legitimacy. It may offer the possibility of transcending national categories such as usual residence by capturing transnational and mobile modes of living (see Chapter 3). However, and in line with the conception developed in this book, data practices are part of a transnational field of statistics where scales of

the local, the national, and the international overlap and intersect and involve complex relations of power and influence such that what they enact are neither 'national' nor 'European' statistics. This is a point which we return to in Chapter 9.

This reflection is critical as digital technologies become ever more part of social life and at the same time part of new data practices for knowing and governing. What we have focused on in this chapter is what this may mean for relations between subjects and the making of population statistics, which are by no means given or settled. Of critical importance is that digital technologies often work in the background: from the technical configurations of digital censuses to the scraping of tweets to infer categories, what then are the possibilities of subversion, intervention, or accountability? Subversion does not only mean to attack or undermine authority but to make democratic demands and claims about its operation. Given the long history of how NSIs have sought to secure the consent of subjects for both the collection and use of data about them, we suggest that possibilities for such democratic interventions and claims are significant, if, as we have argued, being a citizen is to be both a subject to and subject of power, where obedience, submission, and subversion are always-present potentialities. In relation to official statistics, it means to consider subjects as 'data citizens' with the right to shape how data is made about them and the societies of which they are a part, an issue which we return to in the concluding chapter (Ruppert, 2019). That is, the possibilities and potentials of citizens to act in their subjectivation are as important, if not more, than the promises of digital technologies for more timely, efficient, cheaper, and reliable statistics.

References

Adams V, Murphy M, and Clarke AE (2009) Anticipation: Technoscience, Life, Affect, Temporality. *Subjectivity* 28(1): 246–265.

Aradau C and Blanke T (2018) Governing Others: Anomaly and the Algorithmic Subject of Security. *European Journal of International Security* 3(1): 1–21.

Australian Bureau of Statistics (2015) Big Crocs, Big Snakes and Small Censuses: The Story of Australia's Digital-first Census. Paper presented to the Economic Commission for Europe, Conference of European Statisticians. Group of Experts on Population and Housing Censuses, Seventeenth Meeting. Geneva: UNECE.

Balibar E (1991) Citizen Subject. In: Cadava E, Connor P, and Nancy J-L (eds) *Who Comes After the Subject?* London: Routledge, pp. 33–57.

Bexley S (2017) It's All About Inclusion: How ONS Plans to Support the Digital Have-nots. Available at: https://blog.ons.gov.uk (accessed 23 March 2018).

Bruns A and Burgess J (2015) Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research After the Computational Turn. In: Langlois G, Redden J, and Elmer G (eds) *Compromised Data: From Social Media to Big Data*. New York: Bloomsbury Academic, pp. 93–111.

Cakici B and Ruppert E (2019) Methods as Forces of Subjectivation: Experiments in the Remaking of Official Statistics. *Journal of Cultural Economy*: 1–15.

Census Independent Assurance Panel to the Australian Statistician (2017) *Report on the Quality of 2016 Census Data*. Australia: ABS.

Couper MP and Singer E (2013) Informed Consent for Web Paradata Use. *Survey Research Methods* 7(1): 57–67.

Cremonesi L, Irrera O, Lorenzini D, et al. (eds) (2016) *Foucault and the Making of Subjects; Rethinking Autonomy between Subjection and Subjectivation*. London: Rowman & Littlefield International.

Desrosières A (1998) *The Politics of Large Numbers: A History of Statistical Reasoning* (ed. RD Whitley). Cambridge, MA and London: Harvard University Press.

Duke-Williams O (2009) The Geographies of Student Migration in the UK. *Environment and Planning A: Economy and Space* 41(8): 1826–1848.

Foucault M (1982) The Subject and Power. *Critical Inquiry* 8(4): 777–795.

Grommé F, Ruppert E and Cakici B (2018) Data Scientists: A New Faction of the Transnational Field of Statistics. In: Knox H and Nafus D (eds) *Ethnography for a Data Saturated World*. Manchester: Manchester University Press, pp. 33–61.

Guild E (2019) Data Rights: Claiming Privacy Rights Through International Institutions. In: Bigo D, Isin E, and Ruppert E (eds) *Data Politics: Worlds, Subjects, Rights*. Abingdon and New York: Routledge, pp. 266–283.

Haggerty KD and Ericson RV (2000) The Surveillant Assemblage. *British Journal of Sociology* 51(4): 605–622.

HM Government (2018) *Help Shape Our Future: The 2021 Census of Population and Housing in England and Wales*. White Paper. UK Parliament.

Isin E and Ruppert E (2015) *Being Digital Citizens*. London: Rowman & Littlefield International.

Kockelman P (2013) The Anthropology of an Equation: Sieves, Spam Filters, Agentive Algorithms, and Ontologies of Transformation. *HAU: Journal of Ethnographic Theory* 3(3): 33–61.

König R and Rasch M (2014) *Society of the Query Reader: Reflections on Web Search*. Amsterdam: Institute of Network Cultures.

Langlois G, Redden J, and Elmer G (eds) (2015) Introduction. In: *Compromised Data: From Social Media to Big Data*. New York: Bloomsbury Academic, pp. 1–14.

Liu W and Ruths D (2013) What's in a Name? Using First Names as Features for Gender Inference in Twitter. In: *AAAI Spring Symposium: Analyzing Microtext* (conference paper), pp. 10–16.

MacGibbon A (2016) *Review of the Events Surrounding the 2016 eCensus:* Australian Government, Department of the Prime Minister and Cabinet.

Marres N (2017) *Digital Sociology: The Reinvention of Social Research*. London: Wiley.

Mislove A, Lehmann S, Ahn Y-Y, et al. (2011) Understanding the Demographics of Twitter Users. In: *Fifth International AAAI Conference on Weblogs and Social Media* (conference paper).

Mitchell R, Dorling D, Martin D, et al. (2002) Bringing the Missing Million Home: Correcting the 1991 Small Area Statistics for Undercount. *Environment and Planning A: Economy and Space* 34(6): 1021–1035.

Mocanu D, Baronchelli A, Perra N, et al. (2013) The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLOS ONE* 8(4): e61981.

Neyland D and Milyaeva S (2016) The Entangling of Problems, Solutions and Markets: On building a market for privacy. *Science as Culture* 25(3): 305–326.

Nguyen D-P, Gravel R, Trieschnigg RB, et al. (2013) 'How Old Do You Think I Am?' A Study of Language and Age in Twitter. In: *Seventh International AAAI Conference on Weblogs and Social Media* (conference paper).

ONS (2012) 2011 Census: Key Statistics for local authorities in England and Wales. Available at: http://bit.ly/2ncI70x (accessed 28 March 2017).

ONS (2015a) Research for 2021 Census England and Wales: possible innovations under consideration. Paper presented at the Economic Commission for Europe, Conference of European Statisticians. Group of Experts on Population and Housing Censuses. Seventeenth Meeting. Geneva, 30 September to 2 October 2015.

ONS (2015b) *Using Geolocated Twitter Traces to Infer Residence and Mobility*. Available at: http://bit.ly/2mLVFof (accessed 28 March 2017).

Oopkaup A and Servinski M (2013) A Positive View of Demographic Trends. *Eesti Statistika Kvartalikiri [Quarterly Bulletin of Statistics Estonia]* 2013(3): 15–18.

Ratner H (2019) Europeanizing the Danish School through National Testing: Standardized Assessment Scales and the Anticipation of Risky Populations. *Science, Technology, & Human Values*: 1–23.

Ruppert E (2011) Population Objects: Interpassive Subjects. *Sociology* 45(2): 218–233.

Ruppert E (2012) The Governmental Topologies of Database Devices. *Theory, Culture & Society* 29(4–5): 1–21.

Ruppert E (2018) Sociotechnical Imaginaries of Different Data Futures: An Experiment in Citizen Data. *3e Van Doornlezing*, Rotterdam, NL, 14 June. Erasmus School of Behavioural and Social Sciences, Erasmus University Rotterdam.

Ruppert E (2019) Different Data Futures: An Experiment in Citizen Data. *Statistical Journal of the International Association for Official Statistics* 35: 633–641.

Ruppert E and Scheel S (2019) The Politics of Method: Taming the New, Making Data Official. *International Political Sociology* 13(3): 233–252.

Scheel S, Grommé F, Ruppert E, et al. (2016) *Transcending Methodological Nationalism through Transversal Methods? On the Stakes and Challenges of Collaboration*. ARITHMUS Working Paper 1. Available at www.arithmus.eu (accessed 12 May 2018).

Scheel S, Ruppert E, and Ustek-Spilda F (2019) Introduction: Enacting Migration Through Data Practices. *Environment and Planning D: Society and Space* 37(4): 579–588.

Sloan L, Morgan J, Burnap P, et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLOS ONE* 10(3): e0115545.

Statistics Austria (2015) Integrating the Web Mode in the Austrian Household Budget Survey 2014/15. Presentation at Eurostat New Technologies and Techniques in Statistics Conference. Brussels.

Statistics Estonia (2012) *Summary of the Conduct of the Estonian Population and Housing Census (PHC2011)*. Tallinn: Statistics Estonia.

Tiit E-M (2013) Estonian Census 2011. *Papers on Anthropology* (22): 234–246.

Tiit E-M (2015) The First Census Without Enumerators. *EestiPäevaleht*. Available at: http://epl.delfi.ee/archive/article.php?id=72013177.

UK (2017) *Government Transformation Strategy*. Cabinet Office. Available at: https://bityl.co/6HCB (accessed 3 November 2019).

van Dijck J, Poell T and De Waal M (2018) *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press.