The Promise of Computational Text Algorithms for Augmenting Qualitative Analysis in Demography

Margaret Frye
Princeton University

ABSTRACT

This paper explores the advantages of applying computational text analysis to qualitative data in demography. I begin by examining some challenges that demographers are particularly likely to face in analyzing qualitative data—large amounts of data and the difficulty of comparing themes across external categories—and discussing ways that new tools from machine learning and computer science might help to address these challenges. I then describe three applications of text analysis using a set of conversational journals about HIV/AIDS from Malawi. These applications vary in the extent to which computational techniques either supplement or supplant more traditional methods of qualitative data analysis. In the first example, computational techniques are used only for sample selection; in the second, to analyze particular themes over time, and in the third, a computer algorithm is trained to identify latent themes in the text as well as reveal how they vary over time and across individuals.

DIRECT CORRESPONDENCE TO: Margaret Frye, Sociology Department, Wallace Hall, Princeton, NJ, 08540. Email: mfrye@princeton.edu.

This paper will demonstrate the advantages of using computational text analysis for analyzing qualitative data to inform our understanding of population processes. Computational text analysis was first developed at the intersection of computer science and the humanities, and applications of this method have increased in recent years in economics, political science, and sociology. To my knowledge no scholars have yet demonstrated the utility of this set of methods for the field of demography.

I will argue that computational text analysis can help researchers to address a set of unique priorities and concerns that demographers grapple with when analyzing qualitative data. First, due to the macro-level nature of much of demographic inquiry, and the fact that qualitative data are often linked to large-scale surveys or come from newspaper accounts, policy reports, or other media, demographers often face the issue of having too much data to analyze in-depth. Many scholars have solved this problem by taking a "random sample" of the texts and performing a close reading of this subset of the data. Yet this technique is better suited to quantitative analysis than qualitative inquiry. In particular, if one or more themes of interest are present in only a minority of cases, but are substantively important for the project at large, the random sample may contain only a few cases and lead to underdeveloped theories or inaccurate representations. Computational text analysis provides a set of tools through which analysts can select a sub-sample (to later code or read carefully using more traditional qualitative methods) that is thematically relevant to their subject of inquiry, while also gaining an understanding of how common these themes are across the full corpus.

Second, demographers are often interested in how ideas or themes expressed in qualitative data vary over time or space. But in linking external variables such as time or country to the results of qualitative inquiry, researchers often struggle to avoid biasing the results through their very reading of the texts. The same advantages that make qualitative analysis particularly generative of new theories—that it is reflexive, nuanced, and full of rich details—make it difficult to objectively evaluate these conclusions according to an external set of contrasts. Partway through the long and arduous task of reading and coding our data, we develop a hunch of what our argument might be: e.g., that a particular theme may emerge at a particular time, or that one set of concepts is more often discussed by interview respondents of a certain race or ethnicity. From this point forward, future readings are colored by our vision of our future conclusions, and we are more likely to notice examples that confirm our hunch than those that contradict it. From my own experience and conversations with colleagues, I know that we often try to solve this problem by analyzing the data yet again, or by having another researcher analyze a subset of the data. Still, all readings of qualitative data are subject to the same set of subjective biases. These biases, of course, are not unique to demography, but given our proclivity to make comparisons across easily distinguishable categories—year, country, gender, race—this tendency to read a story into our analysis is particularly difficult to avoid in our field.

Computational text analysis does not replace close reading or subjective theorizing, but it can provide a complementary set of tools that can help researchers address these two issues. I demonstrate the unique advantages that computational text analysis has for demography by using data from the Malawian Journals Project (described below). I present three different types of textual analysis using these data. First, I show how one can use an algorithm to *select a sub-sample of data* most likely to be relevant to the topic of inquiry for close reading and coding. Second, I demonstrate how one can construct dictionaries representing themes of interest and examine *trends over time in the prevalence of specific ideas or topics*. And finally, I describe how structural topic modeling can provide an understanding of *the major concepts and themes present in the corpus of data*, and how these themes vary according to a set of variables. For each case, I describe how each type can help to address a research challenge or answer a particular type of research question that demographers routinely encounter.

Data

The Malawian Journals Project, which began in 1999 and is ongoing, was first designed as a qualitative supplement to a longitudinal household survey (Watkins and Swidler 2009). Frustrated by the limitations of the survey format, the principal investigators hired local participant observers ("journalists") to document conversations related to AIDS that took place around them. Journalists were paid $30 US for each completed notebook. Thus far, the project has produced 1,054 80-page notebooks that average 7,500 words in length; most are publicly available online.[i] Typically, each notebook contains three to four separate conversations that the journalist has recorded over a span of a few days to a few months; this sample of 1,054 journals

yielded 4,080 unique "conversational incidents" between 2000 and 2012. (exchanges between two or more people that are bounded by time and space, see Angotti et al. 2014). Appendix Figure 1 shows the distribution of conversational incidents across years.

The journals are an incredibly rich source of information about daily life in rural Malawi, and these texts provide a number of advantages over more conventional sources of data (Watkins and Swidler 2009). They capture people's opinions as expressed organically, rather than in a formal interview setting. Because the journalists select the episodes to write about and the details to include, the data reveal which concepts are most salient and which ideas are conjoined in this cultural context. These data are ideal for understanding meaning-making as a collective process, as they incorporate the disputes, humor, and contradictions that are inherent to everyday talk but typically are missing from sociological data. With a corpus of 4,080 texts, however, it is almost impossible to give each of these texts a close reading, and thus this data source is an ideal candidate for computational text analysis. At the same time, they present unique challenges that many sources of qualitative data used by demographers will likely share: they contain misspellings and words from foreign languages, grammar is often poor, and they often start and stop mid-story. They thus present an excellent test case for whether computational text analysis will be of use to demographers.

Study 1: Using Computational Techniques to Improve Sample Selection

In the first study, a coauthor and I were interested in examining how men evaluate women whom they encounter in public spaces, in terms of their physical attractiveness and their potential HIV risk (a copy of this paper is available from the author upon request). We were interested in conversational incidents involving men in a public setting commenting on particular women whom they saw or interacted with. If we simply analyzed a random sample of the journals, we would have been left with only a few such relevant cases. Instead, we used a combination of traditional qualitative analysis and computational text analysis to select a set of incidents that were particularly relevant to this topic.

Specifically, we used a dictionary-based approach, which begins with a list of words or phrases that are related to a concept or topic of interest, and generates counts of how often each term appears in each text in the corpus, as well overall scores that aggregate across all words included in the dictionary.[1] We relied on two different dictionaries to select our sample. The first dictionary includes words related to attractiveness. To construct a list of words occurring in the journals related to physical attractiveness, we began with a set of 488 conversational incidents that had previously been summarized for another project (Angotti et al. 2014), and used these summaries to identify 81 incidents that seemed to involve commentary on particular women. We both read all incidents in this preliminary sample, and constructed a list of 36 terms that were frequently used by men to describe whether or not women were considered to be attractive (see Appendix Table 1). These correspond to four categories: body parts, fashion or clothing, verbs, and adjectives. The second dictionary identifies incidents in which women are observed in a casual setting. Our preliminary analysis revealed that these incidents typically occur in the market or on the road and involve a description of a woman "passing by" the conversation participants. We thus constructed a dictionary composed of the terms "pass by," "walk by," and "walk past" (including all tenses). By juxtaposing the terms related to attractiveness with those related to women passing by, we were able to identify 180 texts that were topically relevant and also to capitalize on a real strength of our data—its potential to capture discussions sparked by everyday social encounters. We were also able to ensure that our sample was evenly distributed across the 12 years of data, by selecting the entries with the highest ranking for each of several time periods.

In this case, a simple application of computational text analysis enabled us to identify a more salient and informative set of texts than a random sample would have provided, but the analysis itself is based on detailed qualitative coding of all texts in this analytic sample. In other words, this study used computational text analysis only in the initial phase of the research: sample selection.

[1] To normalize across different document lengths, these counts are divided by the total word count for each document.

Study 2: Using Text Analysis to Trace Specific Themes Across Time

In the second study, I again used a dictionary-based approach to ascertain the frequency with which particular words appear in each document. However, this time I focused on the frequencies themselves, and what they can tell us about how discussions related to specific topics around AIDS ebb and flow during the period analyzed in the data. During the period captured by the conversational journals (2000-2012), two technologies emerged that provided rural Malawians with new tools to manage the AIDS epidemic: first, the increasing availability of affordable HIV testing and second, the rollout of antiretroviral therapy (ART). These new technologies rendered AIDS at once more knowable and less visible—people no longer needed to wait until they fell sick to "know their status," yet those who were infected could stay healthy longer without showing symptoms of the disease. Previous research has suggested that these shifts were reflected in the ways that local Malawians discuss and debate about AIDS in the journals (Angotti et al. 2014; Conroy, Yeatman, and Dovel 2013; Watkins 2004). Here, I seek to analyze the entire corpus of journals over time, to determine whether we see a shift in the frequency and intensity with which certain topics (here defined in terms of specific sets of words) are discussed in the journals.

Using the same strategy of reading a preliminary subset of journals to look for relevant terms, I constructed dictionaries representing words related to prevention, testing, and treatment (see Appendix, Table 2). I also constructed a dictionary with words related to death and sickness as well as one with words related to survival and health. I intentionally kept these dictionaries small and focused only on words that I was confident were related to the concepts of interest (in the full version of this paper, I will discuss the reliability checks I conducted on these dictionaries). I then graphed the proportion of all words appearing in the journals over time that relate to each dictionary, using loess smoothing.

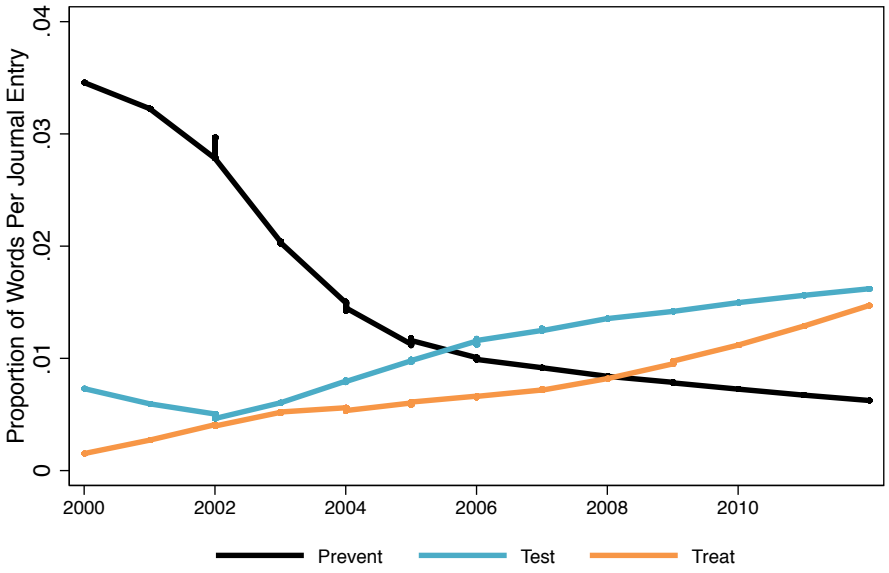
Figure 1 shows the proportion of words over time that fall into the prevention, testing, and treatment dictionaries, respectively. The graph shows a clear trend that is consistent with what we would expect: as testing and treatment technologies became more widely available, the conversations documented in the journals shifted away from abstinence and trust and towards testing. Towards the end of the decade, we also see a rise in the proportion of incidents related to treatment. Figure 2 shows a similar set of results for the "death" versus "life" dictionaries. Here, we see a striking pattern: conversations seem to have shifted over time from being primarily related to death, funerals, and sickness to being about life, survival, and being healthy. In the full version of this paper, I will compare the timing of these shifts in the content of local conversations to changes in testing, ART availability, and AIDS mortality and prevalence rates.

Study 3: Using Topic Modeling to Inductively Explore Themes Over Time and Across Journalists

In the third study, I will use structural topic modeling, a technique that is at the cutting edge of the field of machine learning and text analysis. Topic models allow for rich arrays of "latent" topics to be uncovered from a set of texts. They are often referred to as "unsupervised" because they "*infer* rather than assume the content of the topics under study *(*Grimmer and Stewart 201*3).* The innovation that makes *structural* topic models seem particularly appropriate for demographic analysis is that they incorporate information about the document, such as information about an interviewee based on past interviews, or in the present case, information about the timing of data collection and the journalists who recorded the data (Roberts et al. 2014). This technique has been used successfully across a wide variety of data sources, including newspaper articles, political speeches, tweets, and open-ended survey responses.

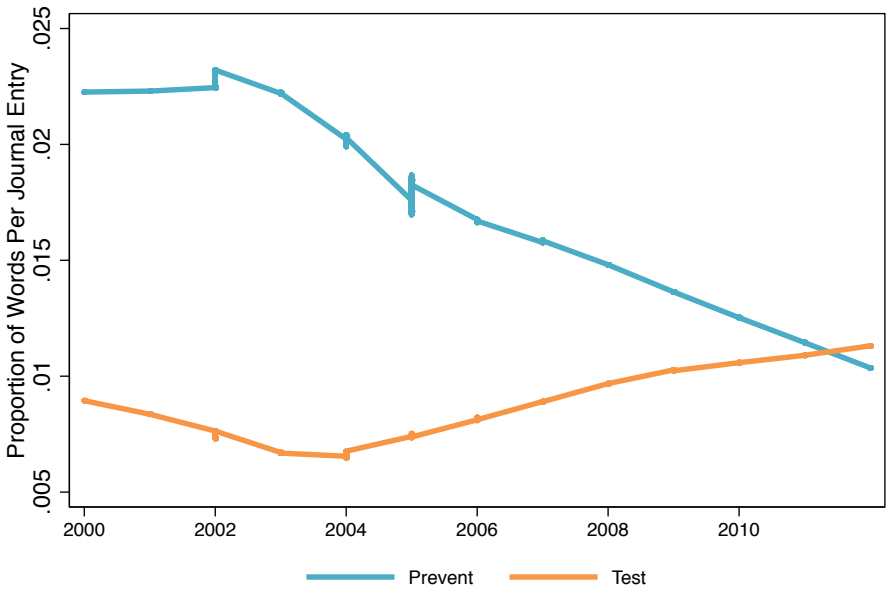
I have completed all of the steps involved in preparing the data to be read into the topic modeling algorithm and have made significant progress towards learning this new method, but have not yet implemented it. It is my hope that by applying this technique I will gain new insights into how the content of local conversations about AIDS has changed over time, through this inductive method's potential to uncover themes that may have been previously overlooked by researchers reading the data with other questions in mind. Of course, it is also possible that the unique nature of the journals—specifically their varying length and non-standard use of English—will make this particular technique less applicable. If this is true, I expect that examining this as a negative case—an example of a computational text method that is not informative when applied to these data—would be helpful in building an understanding of what computational text techniques can and cannot offer to the studies of demographic phenomena.

Figure 1: Proportion of Words Included in "Prevention", "Testing", and "Treatment" Dictionaries



Prevent includes: prevent, abstain, abstinence, faithful, faith, avoid, protect, trust, careful, and condoms.
Test includes: diagnose, test, HTC, status, disclose, VCT, counselor, positive, negative, and result.
Treat includes: antiretroviral, retroviral, pill, tablet, treat, treatment, drug, medicine, ART, ARV, and diet.

Figure 2: Proportion of Words Included in "Death" and "Life" Dictionaries



Death includes: die, death, dead, kill, sick, thin, and funeral.
Life includes: live, life, alive, survive, health, fat, and energy.

References

Angotti, Nicole et al. 2014. "Popular Moralities and Institutional Rationalities in Malawi's Struggle Against AIDS." *Population and Development Review* 40(3):447–73.

Conroy, Amy, Sara Yeatman, and Kathryn Dovel. 2013. "The Social Construction of AIDS during a Time of Evolving Access to Antiretroviral Therapy in Rural Malawi." *Culture, Health & Sexuality* 15(8).

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–97.

Roberts, Margaret E. et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–82.

Watkins, Susan C. 2004. "Navigating the AIDS Epidemic in Rural Malawi." *Population and Development Review* 30(4):673–705.

Appendix

Figure A1: Number of Conversational Incidents in the Corpus for Each Year (Total=4080)

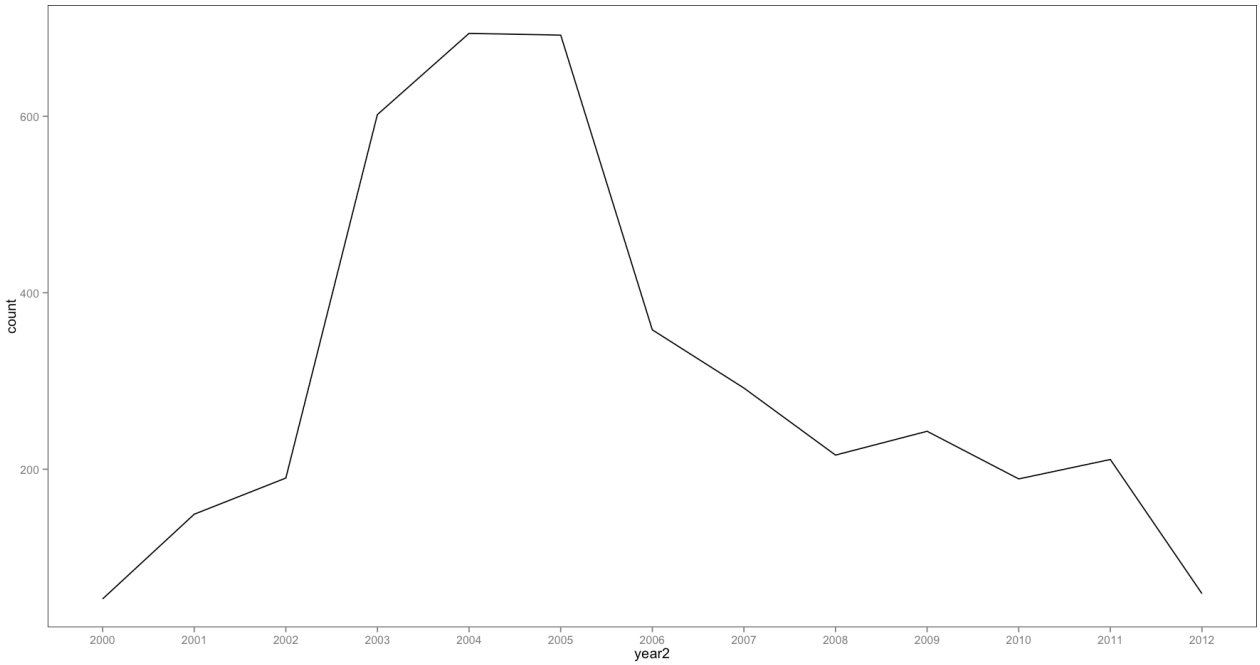


Table A1: Words Included in Attractiveness Dictionary for Study 1

Word Type	Words Included	Average # Included in Sample
Body parts	*Hair, hips, legs, buttocks, butt, thigh, shoulder, back, arms, waist, stomach*	1.6
Clothing	*Naked, pants, fashion, tight, style, skirt, dress, miniskirt, clothes, clothing, blouse, cosmetics*	7.4
Verbs	*Tempt, attract, wear*	2.0
Adjectives	*Beautiful, handsome, sexy, young, pretty, smart, fat, plump, seductive, bare*	5.6

Table A2: Words Included in Dictionaries for Study 2

Dictionary	Words Included
Prevention	Prevent, abstain, abstinence, faithful, faith, avoid, protect, trust, careful, condoms
Testing	Diagnose, test, HTC, VCT, status, disclose, counselor, positive, negative, result
Treatment	Antiretroviral, retroviral, ART, ARV, pill, tablet, treat, treatment, drug, medicine, diet
Death/Sickness	Die, death, dead, kill, sick, thin, funeral.
Life/Health	Life, live, alive, survive, health, fat, energy