Developing Computational Approaches to Investigate Health Inequalities

Kivan Polimis

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Hedwig Lee, Chair

Kyle Crowder, Chair

Emilio Zagheni

Program Authorized to Offer Degree:

Sociology

University of Washington

Abstract

Developing Computational Approaches to Investigate Health Inequalities

Kivan Polimis

Co-Chairs of the Supervisory Committee:

Professor Hedwig Lee

Department of Sociology, University of Washington

Professor Kyle Crowder

Department of Sociology, University of Washington

Computational approaches to health, especially approaches harnessing "big data", offer researchers emerging methods and novel data to understand social inequalities. Through data sources such as social media and smart city technology (e.g., streaming passenger sensor data from public buses), the public is generating trillions of data points and drawing the attention of researchers interested in classic subject areas such as understanding development of inequalities and newer research areas like the connection between online and offline social worlds. Big data has unique strengths and limitations that are magnified by the methods used to analyze big data in the social sciences, a methodology broadly known as computational social science. Computational social science employs computational approaches such as cryptography and machine learning algorithms to analyze, simulate, and model behavioral phenomena. Additionally, computational methods provide large-scale access to unstructured data types such as text, images, and audio that previously eluded social science research. This dissertation explores the substantive and methodological contributions and limitations

of blending big data, computational social science, and synthetic data in two health domains. Synthetic datasets are used to maintain data privacy and simulate data entry uncertainty by leveraging multiple imputation and most importantly for these studies, intentionally inaccurate (e.g., not the actual collected value in the original data set) data. The first study uses Twitter and merges image and geolocation data to assess demographic variations in physical activity attitudes. Chapter one demonstrates how an integrated data gathering and refinement approach can produce a high quality social media data set. Sentiment polarity findings indicate that racial minorities (especially women) discuss physical activity as positive and often more positive than whites challenging conventional hypotheses of race and physical activity attitudes. The second and third chapter use synthetic data based on Homeless Management Information Systems (HMIS) administrative data from a county comprising one major metropolitan area. HMIS data was collected from multiple service providers in federally funded housing programs such as transitional housing, emergency shelters, rapid-rehousing, support services only, etc. The second chapter's substantive aim is to identify family typologies of homelessness service use. Identifying typologies of service use is a goal of homelessness research to align homeless interventions with users and identifying family typologies have become a research focus over the last twenty years. The substantive goal for the third chapter is to examine how family characteristics such as household structure and parents' race interact with housing program interventions to influence homelessness exit pathways. Researchers have disputed how influential housing program and family characteristics are for homelessness duration or persistence in homeless cycles making targeted interventions difficult. The methodological contribution of this chapter is to use random forest classifiers to predict exit pathways across different family characteristic constellations. This dissertation engages multiple computational social science approaches with synthesized administrative and social media data. Strategies and results reveal the potential and pitfalls from leveraging emerging methodology and data to investigate health disparities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am thankful for the mentorship and friendships that have prepared me to be the person I am today. Kyle Crowder, Hedy Lee, Emilio Zagheni, Ariel Rokem, and Daryl Holman have been essential in developing and challenging my ideas to stand on their own. The advising I've received at the University of Washington has been instrumental in my career. Thank you to my committee and all the educators and support staff (looking at you, Savery 211) that have helped me on this journey.

I have also been benefited from encouraging and critical graduate colleagues and friends throughout my academic journey. Having the unique opportunity to be a part of two cohorts and cultures, I've made great personal relationships that have also pushed my work. I want to thank all graduate students in our department generally and Marco Brydolf-Horwitz, Steve Karceski, Chuck Lanfear, Maria Vigna-Loria, Rebecca de buen Kalman, Mike Esposito, Tim Thomas, Yuan Hsiao, Connor Gilroy, and Frank Edwards specifically. My cohort from the University of North Carolina have always been very encouraging. I am glad to have Sarah Gaby and Risa Griffin in my corner. Lastly, Thomas Juarez, Brooke Simonson, Alex von Lockner, Vittal Kaddapakkam, Ran Tang, Erber Hernandez, Alex Okafor, Jeremy Jennings, Daniel Watford, Miguel Vergara, Jeremy Cox, Alex Bernick, Tiffany Johnson, and Peter Capkovic are tremendous friends that have generously volunteered their time and helped me push through hurdles big and small.

The work done in this dissertation has been generously funded and cultivated by multiple organizations. I am grateful to the UW Graduate School, UW Department of Sociology, Microsoft, the Gates Foundation, and the National Science Foundation for providing financial

support for this research. I am also very appreciate of the eScience Institute, an environment that exposed me to many new computational approaches and encouraged my intellectual curiosity with programs such as Data Science for Social Good and Community Data Science Workshops.

Finally, Constance, Mario, Kayla, and my extended family have been tremendously supportive from day 1. There are generations of effort from aunts, uncles, cousins, and a grandmother that have worked to put me in successful positions from before I knew of a day 1. As verbose as this dissertation can be, I am short on words to describe how foundational my family is to me.

DEDICATION

To enquiring minds.

Chapter 1

INTRODUCTION

> "Wittgenstein's ruler: Unless you have confidence in the ruler's reliability, if you use a ruler to measure a table you may also be using the table to measure the ruler." (Taleb 2001, pg. 244)

Computational approaches to health, especially approaches harnessing "big data", offer researchers emerging methods and novel data to understand social inequalities. The infancy of big data and methods in computational social science research makes the field a Wild West frontier of ethical dilemmas and ensemble research strategies from multiple disciplines (Salganik 2017). The recent intersection of computational big data methods and social science research has fittingly left most of the history of computational social science online. For instance, Google Trends data show the meteoric rise of big data and other related terms (big data, machine learning, and computational social science. See Figure 1.1) in Google searches since 2012[1]. Computational social science, a coalescing discipline, is significantly behind the other search terms. The youth of computational social science is also evidenced by recent discipline milestones such as the creation of the first Ph.D. program in 2002[2]

[1]Figure 1.1 queried Google Trends with the search terms big data, machine learning, and computational social science. This query is reproduced here: `https://trends.google.com/trends/explore?date=all&q=big%20data,machine%20learning,computational%20social%20science,data%20science`. Researchers are also turning to Google trends to investigate demographic differences in household dynamics such as fertility behavior (Ojala et al. 2017).

[2]I corresponded with founding faculty from George Mason's Department of Computational Social Science

and the formation of domestic and international conferences three and fifteen years ago, respectively[3].

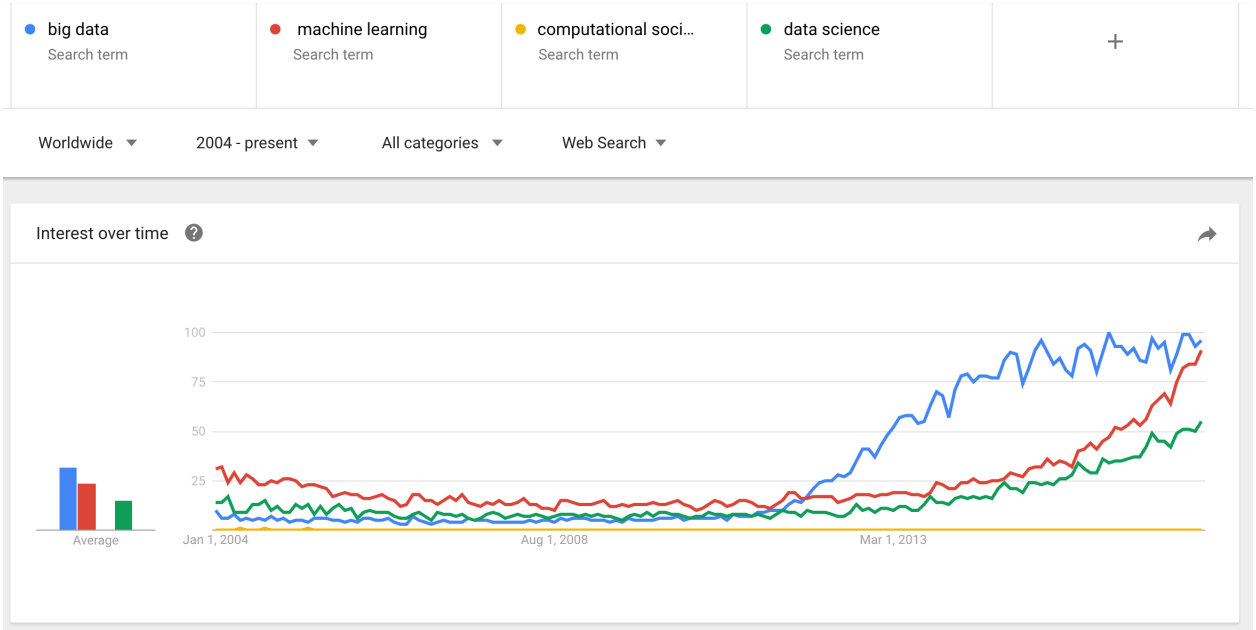


Figure 1.1: Google searches of big data and related terms

Before addressing the possibilities and limitations of big data and computational social science, these amorphous terms that exploded into public and research discussions must be defined. Big data is characterized by the "4 V's" of data – volume, variety, veracity, and velocity[4]. Volume is concerned with the scale and physical size of the data (e.g., terabytes), variety relates to the sources of data (e.g., videos, audio, or text data from social media), veracity addresses data quality and uncertainty from predictions, and velocity corresponds

---

after seeing the claim that their department was the first to offer a Computation Social Science Ph.D. in the following Curriculum Vitae: `http://www.aiecon.org/herbertsimon/series14/CV-Cioff.pdf` These faculty members confirmed that George Mason was indeed the first program, classes didn't start until 2005 and the first graduate completed the program in 2010.

[3]Domestic Conference: `http://www.casos.cs.cmu.edu/naacsos/history.php`
International Conference: `https://ic2s2.org/2017/`

[4]`http://www.ibmbigdatahub.com/infographic/four-vs-big-data`

with the speed data is created (e.g., seconds-based accelerometer data from wearable health devices). Through data sources such as social media and smart city technology (e.g., streaming passenger sensor data from public buses), the public is generating trillions of data points. These individual and meta-data have drawn the attention of researchers interested in classic subject areas such as investigating urbanization and newer research areas like the connection between online and offline social worlds (Athey 2017; Blumenstock, Cadamuro, and On 2015; Ellison, Steinfield, and Lampe 2007; Grieve et al. 2013; Jean et al. 2016; Jerez et al. 2010; Stephens-Davidowitz 2017; Subrahmanyam et al. 2008; Sundsøy 2016). For instance, researchers are using passively collected administrative transit card data and call data records to understand social inequalities. Transit studies have segmented transit ridership by socioeconomic status (SES) to understand equity issues around transportation development.[5] Similarly, behavioral patterns from call data records can be associated with SES and used to predict country-wide poverty distributions. Additionally, social media research suggests that online and offline information dissemination is influenced by demographic inequalities in social media participation and friendship networks (Baeza-Yates 2016; Bozdag 2013; Hajian, Bonchi, and Castillo 2016; Kirkpatrick 2016; Oser, Hooghe, and Marien 2013; Tumasjan et al. 2010).

The ability to passively collect immense amounts of data potentially related to social phenomena is among the many unique strengths and limitations that are magnified by the methods used to analyze big data in the social sciences, a methodology broadly known as computational social science. Returning to the transit card project discussed earlier, a potential shortcoming for researchers centered on data generation limitations and historical inequalities that could produce algorithmic biases. This study's demographic estimates of ridership behaviors countered incomplete coverage of sensor data (not every vehicle in the fleet had passenger counting sensors) and potentially under-counting low-income ridership (these users tend to use cash instead of transit cards) with bias estimations instead of allowing

[5] `https://uwescience.github.io/DSSG2016//2016/07/05/orca-week-3.html`

usage algorithms to replicate historical inequities from limited data. Likewise, Tumasjan et al. (2010) argue for bias corrections that avoid overstating or understating public support for political parties and politicians with sentiment analysis of social media messages because a minority of super users with extreme message volume disproportionately influence overall message polarity.

Algorithms and data are also influenced along multiple demographic features and characteristics of user behavior. Research on potentially discriminatory algorithmic behavior based on demographic features such as gender, ethnicity, or health status is growing (Hajian et al. 2016). Moreover, social media data is likely shaped by network qualities and behavior commercialization. Baeza-Yates (2016) discuss how the activity bias (a minority of users generate most of the content) and presentation bias (past clicking behaviors or ratings changes future interactions with the website/ad delivery) potentially influence social media artifacts. Algorithmic biases and data limitations biases (e.g., historical bias in the data generation process and the unknown impact of online social networks and proprietary business) practices make it difficult to disentangle (raw) data biases and measurement biases. The result of uncertainty is limits on both measurement precision and forecasting (Taleb 2001, 2007, 2012).

The art in computational social science research is managing the trade-offs between methodology and research insight. To this end, computational social science employs computational approaches such as cryptography and machine learning algorithms (e.g., random forests, clustering methods) to analyze, simulate, and model behavioral phenomena[6]. Additionally, computational methods provide large-scale access to unstructured data types such as text, images, and audio that previously eluded social science research[7](Chang, Kauffman, and Kwon 2014; Gandomi and Haider 2015; King 2011). Computational social science as

[6]`https://computationalsocialscience.org/`

[7]Ethnographic social science approaches do incorporate unstructured data, just not at the scale available with computational approaches

a field balances the inherent complexity in combining fields describing social phenomena with tools from the natural and artificial world of engineering and computer science. Social science researchers are broadly asking, "Can new computational methods explore social phenomena?" Fewer researchers[8] are asking, "How can the engineering and computer science tools that represent the backbone of computational social science be modified to capture and investigate social phenomena?"

The asymmetry between problems looking for tools and tools looking for problems is a tension every computational social science study should address. For instance, sentiment analysis has grown in popularity in the social sciences, but the method was originally designed to analyze movie and product reviews[9](Pang and Lee 2005; Pang, Lee, and Vaithyanathan 2002; Turney 2002). Researchers should acknowledge how the method's history (tool) affects their research (problem); reviews, unlike social media, are written with largely informative language and are usually devoid of rhetorical devices such as puns and sarcasm that invert the meaning of information delivered (making overtly positive language a slight and vice-versa). Thus, when sentiment analysis is used for social media data, the method can obfuscate meaning.[10]

Data in this dissertation, like all data generated from an uncertain process, is regarded with caution. Concerns with data biases and computational limitations such as algorithmic bias temper findings. Computational social science is a field for cautious explorations, and findings from this field are one dimension from a broad spectrum of research. Researchers in this area need to exhibit caution when disentangling methods and findings from data

---

[8]Bail (2014) provides a rare framework for integrating theory-driven qualitative cultural sociology computer science methodology.

[9]For an in-depth review of the sentiment analysis and its origins in natural language processing see (Pang and Lee 2008)

[10]Emoticons (emojis) and intensifiers (e.g., all-caps and character repetitions) are also important parts of social media speech that fall outside the sentiment analysis' traditional scope

limitations. Algorithms designed to predict or cluster can exaggerate biases in the data generation process and/or historical legacies of inequality. Although computational methods have brought new data sources under the purview of social science, these data sources and methods are not above scrutiny and should speak to existing research from domain experts. The ability to investigate questions elusive to conventional data sources will develop as computational social science researchers engage with traditional researchers (researchers with more domain expertise and not expertise in using novel data types such as digital trace data/web surveys or machine learning methods). A key step to bridging computational and traditional approaches is the development of methodological clarity and ability to discuss reliability and validity by computational researchers.

1.1 Mo' Data, Mo' Problems?

The greatest strengths and weakness of big data are captured in the name. As the data gets bigger, so does the noise (the ratio may even increase exponentially instead of linearly). Thus while increasing data volume makes it easier to sample groups (especially underrepresented groups), spurious relationships are likely to develop because of the tremendous potential for (non-meaningful) correlations. Moreover, high-dimensional data may produce the "curse of dimensionality", a machine learning term meant to capture prediction problems created from too many predictor variables (Bellman and Dreyfus 1957; Friedman 1997; Keogh and Mueen 2011). Social media research using linked digital trace data (e.g., profiles from multiple social media accounts) are explicitly invoking the curse of dimensionality in their analytical strategies. Tang and Liu (2012) demonstrate how methods to reduce the feature (predictor) space such as supervised learning (user labeled data) outperform kitchen-sink models that include as much information as possible like unsupervised learning (unlabeled data). Skowron et al. (2016) show that unsupervised learning algorithms can be modified to compact feature representation with summary statistic extraction (e.g., mean, standard deviation, etc.) and sub-sampling.

Three key questions should be investigated by computational social science research:

(1) What was measured? (2) What is being modeled? and (3) What are the strengths and limitations of the analytical approach? The nascent stage of big data and computational social science analytics bring reliability and validity concerns to the forefront for researchers and the public alike (Salganik 2017). Although research has demonstrated social media's comparability with offline behaviors in some areas with ground truth data (e.g., voting preferences from social media posts can be verified with voting behavior; see Barberá and Rivero (2015)), ground truth data for other social phenomena (e.g., dating preferences) are more elusive, limiting generalizability. Additionally, numerous unobserved norms can influence interpretations of social media. For instance, gendered or racial communication patterns (clearly in existence offline) could color language use and eventual sentiment analysis conclusions[11].

Contextual questions surround the measurement of social media data in particular. Researchers should ask: "Are individuals using social networks in a way that is measurable for my research question? Is there consistency between behavioral patterns such as attitudes and peer group formation for online and offline persona?" Answers to these and similar questions should be addressed within a social media project. For instance, while individuals might share an unsolicited opinion via Twitter, researchers are unable to observe the rationale (e.g., anger, boredom, attention-seeking behavior, etc.) behind a digital trace observation. However, attention to behavioral patterns such as average sentiment across opinions can inform the signal behind each individual trace. Furthermore, the value in an unsolicited opinion is not how individuals would frame their motivations post hoc, but the fact that they offered an opinion at all. From this perspective, we acknowledge why we might discount some opinions instead of claiming all unsolicited opinions are invalidated. It should be noted that asking individuals their own motivations for actual behaviors after the fact through retrospective recall comes with its own difficulties (De Vera et al. 2010; Falkner, Trevisan,

[11]Sentiment analysis uses computationally intensive techniques to identify positive, neutral or negative opinions in text (Pak and Paroubek 2010)

and McCann 1999; Pettee Gabriel et al. 2011).

1.2 Artificial Intolerance

> "As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic (Caruana et al., 2015; Kim, 2015). Some suggest *model interpretability* as a remedy, but few articulate precisely *what* interpretability means or *why* it is important." (Lipton 2016 pg. 1)

Using big data is akin to searching for a singular needle in a haystack while simultaneously adding more needles and hay to that same stack. Enter, computational social science. While not a panacea, computational approaches grapple head on with the strengths and limitations of big data and these methods are relevant for traditional data. For instance, new computational methods such as facial recognition software can predict an individual's race from an image and rapidly assess demographic characteristics relative to traditional methods. Additionally, statistical methods like decision tree cross-validation can leverage administrative data to build predictions robust to unseen observations with an informed estimate of out-of-sample error. However, the insights gained from computational social science methods should be approached with caution.

Methods are inherently biased. Computational approaches (especially those used to predict) can easily fall victim to algorithmic bias without consideration for historical or data-quality driven biases (Bail 2014; Bozdag 2013; Sweeney 2013). Historically grounded cultural biases are commonly observed with image analysis as well as advertisement and social service delivery. These biases accelerate from offensive to life-altering depending on the algorithmic context. For instance, researchers have shown that image recognition software has labeled black men as apes[12] and image searches of "CEO" under-represent women

[12] https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-

(Kay, Matuszek, and Munson 2015). Algorithms become troubling when they influence ad delivery by suggesting "arrest" to auto-complete Google searches at a very high rate when preceded with names traditionally assigned to Black individuals (e.g., DeShawn, Darnell, and Jermaine) in contrast to names associated more with Whites (e.g., Geoffrey, Jill and, Emma)(Sweeney 2013). These algorithmic-driven biases may be exponentially worse when considering service delivery. Consider the impact of racial biases in software to predict recidivism that is influential in the parole process of several states. Research has documented systematic biases that over-penalize Blacks and underestimate the recidivism of Whites. For instance, the recidivism algorithm consistently overestimated short-term (0-2 year) recidivism rates and upgraded Black sentencing protocols (decreasing parole likelihood) while simultaneously underestimating White prisoners' probability of recidivism and increasing the probability of their parole (Barocas and Selbst 2016; Chouldechova 2017). The very visible, adverse impact on life chances such as re-entering society illustrates the extent algorithms can harm individuals or communities. As a result of data and methodological concerns, studies that leverage big data and computational social science methods should advocate for a cautious analysis of behavioral phenomena.

1.3 Overview

This dissertation explores the substantive and methodological contributions and limitations of blending computational social science and big data in two health domains. The first chapter uses Twitter to assess demographic variations in physical activity attitudes. Physical activity attitudes are related to differences in physical activity engagement, and physical activity is a key predictor of chronic disease development. The substantive goal of Chapter One is to explore whether or not Twitter (big data–volume, veracity, and variety) reveals variations in attitudes towards physical activity by using sentiment analysis to explore opinions expressed towards moderate to vigorous physical activities. Methodologically, this chapter

---

as-gorillas/29567465/

approaches Twitter as a large-scale ethnography of unsolicited opinions and utilizes sentiment analysis, facial recognition software, and geo-located census data (computational methods) to understand demographic differences in physical activity attitudes

The second and third chapter use synthetic data based on Homeless Management Information Systems (HMIS) administrative data (big data–volume, veracity) from a county comprising one major metropolitan area. HMIS data were collected from multiple service providers in Housing and Urban Development (HUD) and Health and Human Services (HHS) funded programs such as transitional housing, emergency shelters, rapid-rehousing, support services only, etc. Each observation represents a unique family's homeless experience: duration, programs used, and (if an exit interview was administered) respondent's living situation at the time of program exit. The data is longitudinal program data from 1993 to 2015 and also includes demographic indicators for each household such as parents' racial background and veteran status. Given HMIS data entry, confidentiality, and privacy concerns, the data used here was synthetically produced from cryptographic methods that alter characteristics of the true data. In the random mode of synthetic data, the data synthesizing software uses random variable sampling, adding the highest degree of differential privacy (noise) (Ping, Stoyanovich, and Howe 2017). The correlated attribute mode uses a Bayesian network to calculate the relationship between variables, and in cases where calculating a Bayesian network is too computationally expensive, Ping et al. (2017) suggest using the independent attribute to sample from a noise-added distribution of the underlying data. Synthetic data further anonymizes individual data and introduces data uncertainty similar to provider (data entry) and respondent (misinformation) inaccuracy observed by homeless service providers using HMIS. Furthermore, models using these data should operate under the uncertainty the data generation process creates. For instance, the combination of political and locational legacies such as the unevenness in the application of selection criteria for a social service program (e.g., disparate program awareness, honest error, or structural racism) affects the composition of populations observed in these programs. Consequently, predictions from these models

should not be generalized to a broader population.

The second chapter's substantive aim is to develop methods that could be used identify family typologies of homelessness service use. Identifying typologies of service use is a goal of homelessness research to align homeless interventions with users; identifying family typologies has become a research focus over the last twenty years (Culhane et al. 2007; Fischer and Breakey 1991; Morse 1992; Shlay and Rossi 1992). The methodological contribution of this chapter is the comparison of clustering strategies across synthetic data sets to create clusters of homelessness service use that can inform family usage typologies and homelessness policy. The clustering strategy relies on finite normal mixture models, models developed to analyze change over time and understand how covariates inform longitudinal processes (Figueiredo and Jain 2002; McLachlan and Peel 2004; Rasmussen 2000).

The substantive goal for the third chapter is to examine how family characteristics such as household structure and parents' racial background interact with housing program interventions to influence homelessness exit pathways. Researchers have disputed how influential housing program and family characteristics are for homelessness duration or persistence in homeless cycles, making targeted interventions difficult (Culhane et al. 2007; Grant et al. 2013 ; Lee, Tyler, and Wright 2010; Metraux and Culhane 1999; Rossi 1991; Shinn 1997; Wood et al. 1990). The methodological contribution of this chapter is to use random forest classifiers (computational methods) to predict exit pathways across different family characteristic constellations. Random forest classifiers are a parametric machine learning method that is robust to noise and outliers. This ensemble technique combines decision trees that individually use input variables to predict group membership for each observation (Breiman 2001a).

These studies explore social science big data research and emerging computational methods to understand variations in physical activity attitudes (social media) and homeless family resource utilization (administrative data). Social media data is newer, and the relationship between social media and the ability to predict "ground-truth" reality is being explored

in domains like poverty prediction (Blumenstock et al. 2015; Jean et al. 2016; Sundsøy 2016). The first chapter on Twitter and physical activity addresses the suspicion with social media data quality and the complexities of porting computer science methodologies to social media data. Analyzing social media data with computational methods like sentiment analysis introduces potential errors from algorithmic bias and incomplete understanding of language richness (e.g., puns, sarcasm, etc.). When analyzing social media data, computational approaches use emerging methods on new data, can magnify the errors based on data generation such as the algorithms behind companies limiting the data available via Application Programming Interfaces (APIs)[13]. This chapter also demonstrates the ability to integrate administrative and facial recognition software and produce estimates of social media users' demographic backgrounds.

Although the administrative data used in the second and third chapter is more familiar to demographic research, the methods used represent emerging approaches being applied by social scientists to understand behavioral phenomena. These chapters are representative of computational methods from computer science and statistics that are traditionally considered atheoretical because they are designed to (1) preserve privacy and (2) maximize predictive performance (Dwork 2006, 2007; Jones and Linder 2015). Social scientists are wary of abstractions from actual data and using "black box" off-the-shelf prediction-oriented methods to understand social phenomena. However, advancements in data cryptography and machine learning interpretability are opening the black box of computational methodology and informing policy by allowing researchers to test multiple hypotheses across potential states of the world.

Lastly, this dissertation hopes to join the movement to increase reproducibility and replicability in the sciences. A 2015 *Nature* survey of 1,576 researchers revealed that 70%

[13]For instance, Twitter's Streaming API allows access to 1% of all tweets occurring in real time without mention of how that 1% of tweets is chosen. Sloan et al. (2013) has shown that the 1% sample is a random representative sample of all tweets]

of researchers have tried and failed to reproduce a study from another researcher (Baker 2016). Additionally, a high-profile controversy surrounding the sociologist Alice Goffman's 2014 book *On the Run*[14], demonstrates that reproducibility challenges exist in computational and traditional research (Lubet 2015). As a nexus between computational and traditional research, this dissertation adopted multiple strategies for transparent research. These strategies include leveraging tools and communities such as version control and Software Carpentry[15] to build data management and computational science skills (Leek and Peng 2015; Stodden 2010). Furthermore, Leek and Peng (2015) suggest that these tools and communities are essential components of reproducible and replicable research that include ensuring "(i) the raw data from the experiment are available [and] (ii) the statistical code and documentation to reproduce the analysis are available" (1645). By using synthetic data, this dissertation is able to share data without violating HMIS data sharing agreements or Twitter's terms of service. All the data and code from this dissertation will be available at my GitHub profile[16] in the hopes of transparency, reproducibility, and replicability.

1.3.1 Summary

> "The trick to being a scientist is to be open to using a wide variety of tools."
> (Breiman 2001b, pg. 214)

The overarching takeaway of this dissertation is to be general and specific with novel data and computational social science. Researchers can be general when re-purposing conventional data such as addresses to improve estimates of social media users race with Census

---

[14] Goffman's response to critics: `http://www.ssc.wisc.edu/soc/faculty/docs/goffman/A%20Reply%20to%20Professor%20Lubet.pdf`

[15] "Software Carpentry is a volunteer organization whose goal is to make scientists more productive, and their work more reliable, by teaching them basic computing skills. Founded in 1998, it runs short, intensive workshops that cover program design, version control, testing, and task automation" from: `https://software-carpentry.org/faq/#what-is-swc`

[16] `https://github.com/kpolimis/`

data. However, researchers need to be specific and avoid out-of-the-box implementations of methods such as sentiment analysis or random forest predictors. The result of a cautious approach is a conscious articulation of the problem inherent in Wittgenstein's ruler and the remedies needed to address data and measurement concerns.

Together, these papers describe a human-in-the-loop approach–a strategy where individuals train algorithms and/or correct inaccurate predictions–to further develop computational social science and health research[17]. This dissertation combines machine efficiency processing large datasets with human awareness of algorithmic and data-curation biases. Allowing humans to influence modeling approaches with theory, real-time, or historical information can clarify computational social science results. This dissertation informs human-driven approaches to social science research by advocating a cautious application of computational methods, analyzing multiple data sources (administrative data and social media), and applying emerging methods to understand health inequalities.

[17]Thanks to Ridhi Kashyap, Allie Morgan, Adaner Usmani and the Summer Institute in Computation Social Science (SICSS) for introducing me to the term "human-in-the-loop"

1.3.2 References

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355(6324):483–85. Retrieved July 20, 2017 (http://www.sciencemag.org/lookup/doi/10.1126/science.aal4321).

Baeza-Yates, Ricardo. 2016. "Data and Algorithmic Bias in the Web." Pp. 1–1 in. ACM Press. Retrieved June 8, 2017 (http://dl.acm.org/citation.cfm?doid=2908131.2908135).

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3-4):465–82. Retrieved August 10, 2017 (http://link.springer.com/10.1007/s11186-014-9216-5).

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533(7604):452–54. Retrieved August 10, 2017 (http://www.nature.com/doifinder/10.1038/533452a).

Barberá, Pablo and Gonzalo Rivero. 2015. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33(6):712–29.

Barocas, Solon and Andrew D. Selbst. 2016. *Big Data's Disparate Impact*. clr. Retrieved (https://pdfs.semanticscholar.org/1d17/4f0e3c391368d0f3384a144a6c7487f2a143.pdf).

Bellman, Richard Ernest and Stuart Dreyfus. 1957. *Dynamic Programming*. 1. Princeton Landmarks in Mathematics ed., with a new introduction. Princeton, NJ: Princeton University Press.

Blumenstock, J., G. Cadamuro, and R. On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264):1073–6. Retrieved July 20, 2017 (http://www.sciencemag.org/cgi/doi/10.1126/science.aac4420).

Bozdag, Engin. 2013. "Bias in Algorithmic Filtering and Personalization." *Ethics and Information Technology* 15(3):209–27. Retrieved June 8, 2017 (http://link.springer.

com/10.1007/s10676-013-9321-6).

Breiman, Leo. 2001a. "Random Forests." *Machine learning* 45(1):5–32.

Breiman, Leo. 2001b. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16(3):199–231. Retrieved May 27, 2017 (http://projecteuclid.org/euclid.ss/1009213726).

Chang, Ray M., Robert J. Kauffman, and YoungOk Kwon. 2014. "Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data." *Decision Support Systems* 63:67–80. Retrieved May 26, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0167923613002212).

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *arXiv preprint arXiv:1703.00056*.

Culhane, Dennis P., Stephen Metraux, Jung Min Park, Maryanne Schretzman, and Jesse Valente. 2007. "Testing a Typology of Family Homelessness Based on Patterns of Public Shelter Utilization in Four U.S. Jurisdictions: Implications for Policy and Program Planning." *Housing Policy Debate* 18(1):1–28. Retrieved January 30, 2017 (http://www.tandfonline.com/doi/abs/10.1080/10511482.2007.9521591).

De Vera, M. A., C. Ratzlaff, P. Doerfling, and J. Kopec. 2010. "Reliability and Validity of an Internet-Based Questionnaire Measuring Lifetime Physical Activity." *American Journal of Epidemiology* 172(10):1190–8. Retrieved May 26, 2017 (https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwq273).

Dwork, Cynthia. 2006. "Differential Privacy." Pp. 1–12 in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, vol. 4052. Venice, Italy: Springer Verlag. Retrieved (https://www.microsoft.com/en-us/research/publication/differential-privacy/).

Dwork, Cynthia. 2007. "Ask a Better Question, Get a Better Answer A New Approach

to Private Data Analysis." Pp. 18–27 in *11th International Conference on Database Theory (ICDT 2007)*, vol. 4353. Barcelona, Spain: Springer. Retrieved (https://www.microsoft.com/en-us/research/publication/ask-a-better-question-get-a-better-answer-a-

Ellison, Nicole B., Charles Steinfield, and Cliff Lampe. 2007. "The Benefits of Facebook 'Friends:' Social Capital and College Students' Use of Online Social Network Sites." *Journal of Computer-Mediated Communication* 12(4):1143–68. Retrieved July 20, 2017 (http://doi.wiley.com/10.1111/j.1083-6101.2007.00367.x).

Falkner, Karen L., Maurizio Trevisan, and Susan E. McCann. 1999. "Reliability of Recall of Physical Activity in the Distant Past." *American journal of epidemiology* 150(2):195–205.

Figueiredo, M.A.T. and A.K. Jain. 2002. "Unsupervised Learning of Finite Mixture Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(3):381–96. Retrieved July 3, 2017 (http://ieeexplore.ieee.org/document/990138/).

Fischer, Pamela J. and William R. Breakey. 1991. "The Epidemiology of Alcohol, Drug, and Mental Disorders Among Homeless Persons." *American Psychologist* 46(11):1115–28. Retrieved April 5, 2017 (http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.46.11.1115).

Friedman, Jerome H. 1997. "On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality." *Data Mining and Knowledge Discovery* 1(1):55–77. Retrieved (http://dx.doi.org/10.1023/A:1009778005914).

Gandomi, Amir and Murtaza Haider. 2015. "Beyond the Hype: Big Data Concepts, Methods, and Analytics." *International Journal of Information Management* 35(2):137–44. Retrieved May 26, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0268401214001066).

Grant, Roy, Delaney Gracy, Grifin Goldsmith, Alan Shapiro, and Irwin E. Redlener. 2013. "Twenty-Five Years of Child and Family Homelessness: Where Are We Now?" *American Journal of Public Health* 103(S2):e1–e10. Retrieved April 8, 2017 (http://ajph.

aphapublications.org/doi/10.2105/AJPH.2013.301618).

Grieve, Rachel, Michaelle Indian, Kate Witteveen, G. Anne Tolan, and Jessica Marrington. 2013. "Face to Face or Facebook: Can Social Connectedness Be Derived Online?" *Computers in Human Behavior* 29(3):604–9. Retrieved July 20, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0747563212003226).

Hajian, Sara, Francesco Bonchi, and Carlos Castillo. 2016. "Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining." Pp. 2125–6 in. ACM Press. Retrieved June 8, 2017 (http://dl.acm.org/citation.cfm?doid=2939672.2945386).

Jean, N. et al. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353(6301):790–94. Retrieved July 20, 2017 (http://www.sciencemag.org/cgi/doi/10.1126/science.aaf7894).

Jerez, José M. et al. 2010. "Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem." *Artificial Intelligence in Medicine* 50(2):105–15. Retrieved July 20, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0933365710000679).

Jones, Zachary and Fridolin Linder. 2015. "Exploratory Data Analysis Using Random Forests." in *Prepared for the 73rd annual MPSA conference*.

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. 2015. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." Pp. 3819–28 in. ACM Press. Retrieved May 19, 2017 (http://dl.acm.org/citation.cfm?doid=2702123.2702520).

Keogh, Eamonn and Abdullah Mueen. 2011. "Curse of Dimensionality." Pp. 257–58 in *Encyclopedia of Machine Learning*. Springer.

King, G. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331(6018):719–21. Retrieved May 26, 2017 (http://www.sciencemag.org/cgi/doi/10.1126/science.

1197872).

Kirkpatrick, Keith. 2016. "Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly?" *Communications of the ACM* 59(10):16–17. Retrieved July 20, 2017 (http://dl.acm.org/citation.cfm?doid=3001840.2983270).

Lee, Barrett A., Kimberly A. Tyler, and James D. Wright. 2010. "The New Homelessness Revisited." *Annual Review of Sociology* 36(1):501–21. Retrieved April 8, 2017 (http://www.annualreviews.org/doi/10.1146/annurev-soc-070308-115940).

Leek, Jeffrey T. and Roger D. Peng. 2015. "Opinion: Reproducible Research Can Still Be Wrong: Adopting a Prevention Approach." *Proceedings of the National Academy of Sciences* 112(6):1645–6. Retrieved August 10, 2017 (http://www.pnas.org/lookup/doi/10.1073/pnas.1421412111).

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *arXiv preprint arXiv:1606.03490*.

Lubet, Steven. 2015. *Ethics on the Run*. Rochester, NY: Social Science Research Network. Retrieved August 13, 2017 (https://papers.ssrn.com/abstract=2611742).

McLachlan, Geoffrey and David Peel. 2004. *Finite Mixture Models*. John Wiley & Sons.

Metraux, Stephen and Dennis P. Culhane. 1999. "Family Dynamics, Housing, and Recurring Homelessness Among Women in New York City Homeless Shelters." *Journal of family issues* 20(3):371–96.

Morse, Gary A. 1992. "Causes of Homelessness." Pp. 3–17 in *Homelessness: A National Perspective*, edited by M. J. Robertson and M. Greenblatt. Boston, MA: Springer US. Retrieved (http://dx.doi.org/10.1007/978-1-4899-0679-3_1).

Ojala, Jussi, Emilio Zagheni, Francesco C. Billari, and Ingmar Weber. 2017. "Fertility and Its Meaning: Evidence from Search Behavior." *arXiv preprint arXiv:1703.03935*.

Oser, Jennifer, Marc Hooghe, and Sofie Marien. 2013. "Is Online Participation Distinct

from Offline Participation? A Latent Class Analysis of Participation Types and Their Stratification." *Political Research Quarterly* 66(1):91–101. Retrieved May 25, 2017 (http://journals.sagepub.com/doi/10.1177/1065912912436695).

Pak, Alexander and Patrick Paroubek. 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." in *Proceedings of the International Conference on Language Resources and Evaluation*. Valleta, Malta.

Pang, Bo and Lillian Lee. 2005. "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales." Pp. 115–24 in *Proceedings of ACL*.

Pang, Bo and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2(1–2):1–135. Retrieved November 7, 2016 (http://www.nowpublishers.com/article/Details/INR-011).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment Classification Using Machine Learning Techniques." Pp. 79–86 in *Proceedings of EMNLP*.

Pettee Gabriel, Kelley, James J. McClain, Kendra K. Schmid, Kristi L. Storti, and Barbara E. Ainsworth. 2011. "Reliability and Convergent Validity of the Past-Week Modifiable Activity Questionnaire." *Public Health Nutrition* 14(03):435–42. Retrieved May 26, 2017 (http://www.journals.cambridge.org/abstract_S1368980010002612).

Ping, Haoyue, Julia Stoyanovich, and Bill Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." P. 42 in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Rasmussen, Carl Edward. 2000. "The Infinite Gaussian Mixture Model." Pp. 554–60 in *Advances in neural information processing systems*.

Rossi, Peter H. 1991. *Down and Out in America: The Origins of Homelessness*. University

of Chicago Press.

Salganik, Matthew J. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Shinn, Marybeth. 1997. "Family Homelessness: State or Trait?" *American Journal of Community Psychology* 25(6):755–69. Retrieved April 8, 2017 (http://doi.wiley.com/10.1023/A:1022209028188).

Shlay, Anne B. and Peter H. Rossi. 1992. "Social Science Research and Contemporary Studies of Homelessness." *Annual Review of Sociology* 18(1):129–60. Retrieved May 9, 2017 (http://www.annualreviews.org/doi/10.1146/annurev.so.18.080192.001021).

Skowron, Marcin, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. "Fusing Social Media Cues: Personality Prediction from Twitter and Instagram." Pp. 107–8 in. ACM Press. Retrieved May 30, 2017 (http://dl.acm.org/citation.cfm?doid=2872518.2889368).

Sloan, Luke et al. 2013. "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter." *Sociological Research Online* 18(3). Retrieved February 21, 2017 (http://www.socresonline.org.uk/18/3/7.html).

Stephens-Davidowitz, Seth. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: HarperCollins.

Stodden, Victoria. 2010. "Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science." *Computing in Science & Engineering* 12(5):8–13. Retrieved August 10, 2017 (http://ieeexplore.ieee.org/document/5562471/).

Subrahmanyam, Kaveri, Stephanie M. Reich, Natalia Waechter, and Guadalupe Espinoza. 2008. "Online and Offline Social Networks: Use of Social Networking Sites by Emerging Adults." *Journal of Applied Developmental Psychology* 29(6):420–33. Retrieved July 20,

2017 (http://linkinghub.elsevier.com/retrieve/pii/S0193397308000713).

Sundsøy, P\a al. 2016. "Can Mobile Usage Predict Illiteracy in a Developing Country?" *arXiv preprint arXiv:1607.01337*.

Sweeney, Latanya. 2013. "Discrimination in Online Ad Delivery." *Queue* 11(3):10.

Taleb, Nassim. 2001. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House Incorporated.

Taleb, Nassim Nicholas. 2007. *The Black Swan: The Impact of the Highly Improbable*. 1st ed. New York: Random House.

Taleb, Nassim Nicholas. 2012. *Antifragile: Things That Gain from Disorder*. Random House.

Tang, Jiliang and Huan Liu. 2012. "Unsupervised Feature Selection for Linked Social Media Data." P. 904 in. ACM Press. Retrieved May 30, 2017 (http://dl.acm.org/citation.cfm?doid=2339530.2339673).

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." *ICWSM* 10(1):178–85.

Turney, Peter D. 2002. "Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews." Pp. 417–24 in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Wood, D., R. B. Valdez, T. Hayashi, and A. Shen. 1990. "Homeless and Housed Families in Los Angeles: A Study Comparing Demographic, Economic, and Family Function Characteristics." *American Journal of Public Health* 80(9):1049–52. Retrieved April 8, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.80.9.1049).

Chapter 2

CAN SOCIAL MEDIA BE USED TO ASSESS DEMOGRAPHIC VARIATIONS IN PHYSICAL ACTIVITY ATTITUDES?

2.1 Introduction

Social media create novel, user-generated forums that can mirror observing human behavior. Attitudinal researchers gravitate towards social media data because these data are unsolicited and produced at a tremendous scale. However, social media data are not without their own limitations and the efficacy of these data to explore social phenomena is continuously being evaluated (Bail 2014). Social media research could provide insight into racial differences in chronic disease prevalence by leveraging individual expression to better understand proximate disease risk factors such as attitudes.

Researchers consistently observe large and persistent racial differences in chronic disease prevalence (Smedley, Stith, and Nelson 2003). Racial minorities have higher comorbidity rates and are susceptible to increased mortality (Cossrow and Falkner 2004). Longitudinal and cross-sectional population health surveys regularly highlight racial and ethnic disparities in protective health behaviors including moderate to vigorous physical activity which may drive disease prevalence disparities (Crespo et al. 2000; Dietz 1998; Schwarz and Peterson 2010; Stephens, Jacobs Jr., and White 1985; Tucker, Welk, and Beyler 2011). Gender is also correlated with physical activity and interacts with race in important ways; for example, minority (Black and Hispanic) women are on average less physically active than white women and Black and Hispanic men (Wilcox et al. 2000).

Although physical activity is determined by individual, social and other ecological determinants, a significant body of research asserts that individual attitudes are impor-

tant physical activity predictors (Ajzen 1991; Azjen 1985; Godin et al. 1987; Hagger et al. 2003). Attitudes influence physical activity independent of many social and ecological factors known to affect physical activity rates such as social networks and built environments (Brenes, Strube, and Storandt 1998; Courneya et al. 2000). As mentioned before, physical activity is lower for racial minorities, particularly minority women. The available attitudinal survey research with diverse samples includes findings that racial minorities often have similarly positive or more positive attitudes towards physical activity than their White peers complicates the established relationship between attitudes and engagement. Moreover, female-specific surveys have found that Black and Hispanic women report more positive attitudes towards physical activity than White women despite health literature showing these women engage in less physical activity than their White peers (Crespo et al. 2000; Eyler et al. 2002; Im, Chang, et al. 2012). Racial similarities in physical activity attitudes despite behavioral differences in physical activity trends challenge well established links between attitudes and behaviors. Either attitudes are less salient for physical activity or current methods have not revealed physical activity attitudinal complexity for minority groups.

The multidimensionality inherent in attitudes suggests survey data may not fully capture attitudinal differences, especially for physical activity. Ethnographic studies examining physical activity attitudes utilizing diverse racial samples implicate barriers, knowledge gaps, activity preferences and other difficult to measure factors that may be related to racial variations in physical activity attitudes (Lavizzo-Mourey et al. 2001). Through open-ended and semi-structured approaches, ethnographies have added nuance to attitudinal variation research by providing a forum for individuals to voice perceived physical activity benefits and constraints. However, ethnographic studies face challenges related to small sample size and limited geographies. A more systematic investigation of racial and gender differences that can both capture attitudinal complexity and address larger population segments is needed to better provide nuance to physical activity attitudes.

This paper's purpose is twofold: to substantively assesses demographic variation in physical activity attitudes and methodologically explore social media data challenges and limitations. First, this paper examines physical activity attitude variation by race and gender with Twitter as a large scale ethnography with respondents' unsolicited views toward various physical activities. Analyses also investigate variations at the intersections of demographic characteristics. Textual data on moderate to vigorous leisure time physical activity from Twitter is analyzed with sentiment analysis to understand demographic variations in physical activity attitudes. Twitter users' demographic background is estimated with the combination of facial recognition software and a Census-based last name by location predictor. These analyses reveal attitudes towards physical activity vary by activity and across racial groups with minimal gender variation within racial groups. Lastly, this study contributes to growing literature on social media and demographic research by highlighting methodological challenges facing Twitter-based demographic studies and providing suggestions to address potential social media data biases. Approaches to uncertainty of measurement across attitudes and demographics create a sensitivity-based approach that is instructive for disentangling results and providing a template for future research.

2.2 Background

This review examines key relationships between physical activity and attitudes, demographic variation in physical activity attitudes, and the viability of social media data to understand physical activity.

2.2.1 Physical Activity and Attitudes

Physical activity is broadly defined as "any bodily movement produced by skeletal muscles that results in energy expenditure" (Pate et al. 1995). Health professionals recommend moderate and vigorous physical activity, measured by metabolic equivalent (MET), because these activities provide substantive contributions to individuals' total caloric expenditure and overall health status (Haskell et al. 2007; Hendelman et al. 2000; Westerterp

and Plasqui 2004). Typically, vigorous activities like running exert greater than 6 METs while moderate activities including walking are equivalent to 4-6 METs (Lee and Paffenbarger 2000).

Numerous models exist for describing the relationship between health behaviors, specifically physical activity, and attitudes. Previous health research has focused on latent measures and adapted psychological constructs to understand factors that affect the attitude-behavior relationship (Ajzen 1991; Ajzen and Timko 1986; Azjen 1985; Giles-Corti and Donovan 2002; Godin et al. 1987; Hagger et al. 2003; Voas 2014). Attitudinal health studies find that attitudes influence key physical activity predictors, including persistently engaging in physical activity and behavioral intention[1]. Individuals with positive attitudes towards physical activity intend to (and measurements confirm) engage in physical activity more regularly and across the lifespan than individuals without these attitudes (Affuso et al. 2011; Hagger, Chatzisarantis, and Biddle 2002; Tammelin et al. 2003). Some researchers have critiqued attitudes' importance to physical activity; however, psychological constructs appear important predictors for health behaviors (see: Trost et al. (2002) for critiques of physical activity attitudes) .

2.2.2 Physical Activity Attitudes and Demographic Variation

Health attitudes (e.g, orientation to physical activity) vary by race and gender as well as the intersection of these demographic backgrounds (Clark and Nothwehr 1999; Courtenay, McCreary, and Merighi 2002; Mcguire et al. 2002). While we have a broad understanding of demographic differences in physical activity trends, we have a limited understanding of attitudinal variation along demographic background intersections (Harden 2004). Exploring demographic characteristics individually and in combination can clarify physical activity attitudinal variation.

---

[1] "Intentions are assumed to capture the motivational factors that influence a behavior; they are indications of how hard people are willing to try, of how much of an effort they are planning to exert, in order to perform the behavior" (Ajzen (1991), pg.181).

2.2.3 Racial Variation

Attitudinal research with multiple racial groups produces conflicting comparisons that highlight differences between survey and ethnographic methods. For instance, Affuso et al. (2011) use a telephone survey with general questions about exercise (Appendix A.1) and find majority agreement amongst African-American men and women agree that physical activity is important. Contrastingly, ethnographic research observes a broader range of minority attitudes towards physical activity influenced by cultural ideals towards self-rated health, body-size, and fatalism (Baptiste-Roberts et al. 2007; Egede and Bonadonna 2003; Krause and Jay 1994). Ethnographic studies suggest that less positive attitudes toward physical activity may reflect cultural norms interacting with ecological constraints. Additionally, attitudinal studies often focus on Black-White differences, limiting information about other minority attitudes towards physical activity, such as Hispanic and Asian-American attitudes. Studies that do investigate Hispanic or Asian physical activity attitudes favor examining acculturation and immigration processes instead of broadly studying these communities (Johnson 2000; Kandula and Lauderdale 2005; Unger et al. 2004)[2]. Acculturation- or immigration-based studies find that the migration experience adversely impacted immigrant health by increasing obesity-related behaviors.

2.2.4 Gender Variation

Gendered social and cultural norms could produce variation in physical activity attitudes. For instance, Eyler et al. (1998) and Dwyer et al. (2006) find that women and girls' attitudes towards physical activity are influenced by gender norms that deter physical activity. Moreover, Hayes, Crocker, and Kowalski (1999) find that women subjectively rated their physical activity engagement self-perceptions lower than men despite engaging in the same objective levels of physical activity or having similar levels of self-esteem. Intersectional studies provide opportunities to assess cultural norm influences on gendered

[2]see (Eyler et al. 1998; Im, Chang, et al. 2012; Im et al. 2008, 2015 for exceptions, 2013, 2010; Im, Y. Ko, et al. 2012).

attitudes towards physical activity. These studies demonstrate how overlapping social identities (e.g., race, gender, age, class, religion, etc.) interact to produce and exacerbate social inequalities (Crenshaw 1991).

Intersectional race and gender studies provide opportunities to understand how cultural norms and ecological dynamics are operationalized into physical activity attitudes but conflicting findings have emerged. For instance, Wilcox et al. (2000) and Grieser et al. (2006) rely on survey scales to conclude that Blacks and Whites have more attitudinal commonalities than differences and caution against race-specific health interventions for males or females. Wilcox et al. (2000) finds that Black and White women endorse exercise for health and a desire to increase current physical activity at similar rates while Grieser et al. (2006) states that "girls from all groups have similar perceptions of the benefits of physical activity, with staying in shape as the most important" (pg. 40). Contrastingly, an internet based midlife women's physical activity attitude study find racial differences in physical activity attitudes with scaled instruments. Im, Chang, et al. (2012) shows that midlife racial/ethnic minority women (Hispanic and Non-Hispanic African American women) report significantly *greater* positive attitudes towards physical activity than Non-Hispanic White women.

Studies using focus groups, semi-structured interviews, and other interactive forums indicate layered complexity behind minority men and women's physical activity attitudes (Airhihenbuwa et al. 1995; Henderson and Ainsworth 2003; Im, Y. Ko, et al. 2012; Versey 2014). Multiple authors suggest that African-American women's less positive physical activity attitudes are influenced by marginalized experiences and cultural beauty norms (e.g., hair maintenance) instead of outright dislike for physical activity. However, because few large-scale studies offer detailed attitudinal questions there is less certainty in generalizing race-specific attitudes. Recent research by Ray (2017) has also shown that intersectionality may be conditioned by neighborhood experience. In a physical activity study of middle class blacks and whites to examine physical activity, Ray (2017) found

that urbanicity and percentage of white neighbors differentially affected black men and women's exercise behaviors. Physical activity attitudinal variation by racial, gender and intersectional identities reveal methodology and item-specificity may also influence results. For instance, studies with survey methods tend to focus on how much attitudes differ by race, while ethnographies investigate why attitudinal differences may exist. To these ends, surveys often measure attitudes with scaled responses from generic survey items about exercise. The emerging attitudinal differences (or lack thereof) from surveys may be artifacts of how individuals self-referentially interpret questions about physical activity[3]. Thus, probing attitudes with user-driven responses related to specific physical activities could enhance our understanding gender variations across physical activity.

2.2.5 Understanding Health Behaviors with Unstructured Data

User-generated, unstructured data provide opportunities to address the subjectivity inherent in asking individuals to declare attitudes towards physical activity via survey or in the presence of a researcher and or peers using interview or focus group methods. Examples of unstructured data include textual, visual and auditory data sources, dimensionally rich information not typically available in traditional survey methodology. The similarity across these data types is the lack of predefined model such that data are not "table-orientated as in a relation model or sorted-graph as in an object database" making it difficult to process with traditional software programs (Abiteboul 1997). This paper utilizes one form of unstructured data–those derived from digital traces or records of online interactions. More specifically, this study leverages Twitter data, typically text-laden

---

[3]Krause and Jay (1994) use in-depth one-on-one interviews that blended survey items with follow-up opportunities to elaborate on how individual reference points (e.g., focusing on health problems versus physical function) influence racial attitudinal variation in self-rated health responses. Their data from open-ended responses suggested that global self-rated health questions are being interpreted differentially by race further demonstrating the utility of ethnographic approaches. Similarly, Boyington, Howard, and Holmes (2008) also finds physical activity criteria for evaluation, reference points such as self-rated health and physical functional limitations, vary by race.

descriptions by individuals describing daily activities or social events (restricted to 140 characters via the online platform). However, increased dimensionality is not without its own pitfalls (Bellman and Dreyfus 1957; Friedman 1997). Digital trace research is demonstrating how explicit methods to reduce potential predictors (features) improve inference quality. Through supervised feature selection (user labeled data) or advanced unsupervised (non-labeled data) learning algorithms that are modified to compact features, social media researchers can reduce uncertainty about network and individual characteristics (Skowron et al. 2016; Tang and Liu 2012).

The discrepancy in attitudinal findings from qualitative and survey research discussed earlier could be related to the subjectivity inherent in attitudes. Popay, Rogers, and Williams (1998) suggest standards for using qualitative research for understanding health attitudes that include data reflecting "interpretation of subjective meaning, description of social context and attention to lay knowledge" (pg. 341). Ethnographers rely on forums that produce user-generated, unstructured data and allow respondents bottom-up attitude descriptions instead of top-down criteria limiting response types. Ethnographic studies provide observational richness on attitudes that surpasses survey data, but ethnographers are limited by survey size and methods to manage potential data biases. Multiple ethnographic studies have generated insight into attitudes from focus groups and semi-structured interviews (French et al. 2005; Mabry et al. 2003; Siddiqi, Tiro, and Shuval 2011) by allowing respondents to drive their response narrative. Ethnography can leverage individual subjectivity to improve survey scale measurement reliability (Krause and Jay 1994). However, these studies have focused on small communities and lack the respondent diversity (survey size and demographic variation).

Unstructured data from digital traces (e.g., textual Twitter data) presents unique advantages and disadvantages when compared with traditional survey instruments and ethnographic studies. Traditional health survey instruments leave individuals with a limited response ranges (e.g., Likert scales) and can be uncertain in their core attitude

measurement (Streiner, Norman, and Cairney 2015). Alternatively, digital data can be gathered inexpensively, rely on user-driven responses and are generated more frequently. New findings comparing social media data to traditional data sources reveals that social media can reflect the ground-truth reality of economic disadvantage and demographic distribution for studies analyzing physical activity, nutrition, and well-being (Nguyen et al. 2016). For instance, research has shown that census-level indicators including economic disadvantage predict less frequent physical activity references for Twitter users residing in those areas. Twitter resembles traditional ethnographic approaches by providing a forum for individuals to freely discuss personal opinions eliciting more respondent control in describing attitudes that is common to ethnographic research. Despite these relative strengths compared to traditional research methodology, social media data projects have unique reliability and generalization concerns.

Profile and audience curation typify reliability concerns with social media data. Studies using social media data grapple with classical sociological concepts such as presentation of self, impression management, and self disclosure that may contextualize social interactions and individual behavior (Bullingham and Vasconcelos 2013; Goffman and others 1978; Hogan 2010; Krämer and Winter 2008). Social media users can digitally "curate" an online persona through word choice, picture selection, and network self-selection that can distort online interactions (Arseniev-Koehler et al. 2016; Kaplan and Haenlein 2010; Papacharissi 2012). The potential to turn online interactions into a self-evaluation prism has lead social media researchers to consider audiences and visible within-person changes[4] to contextualize digital traces. Researchers have recognized demographic differences across social media sites that drive generalizability concerns (Potts and Jones 2011). Among all adult internet users, Twitter is over-represented by young adults and racial minorities (Smith and Brenner 2012). Additionally, Twitter is used by a smaller share of adult

[4]Social media users can always delete accounts, create new profiles, or maintain multiple profiles in ways that make tracking within-person changes difficult.

internet users than other social media sites (e.g., Facebook)(Smith and Brenner 2012). See Cesare et al. (2016) and Müller et al. (2016) for an in-depth review of demographic research using digital traces and big data analytics.

2.2.6 Twitter and Health/Physical Activity Studies

Emerging literature has used social media and digital traces to examine physical activity differences (Cavallo et al. 2012). This literature relies on numerous studies documenting the ability of social media in general, and Twitter in particular, to provide individual and population-health insights through observational study of human behavior (Hawn 2009; McCormick et al. 2015; Nguyen et al. 2016; Paul and Dredze 2011; Scanfeld, Scanfeld, and Larson 2010). Studies examining physical activity using social media have revealed that these data sources provide opportunities to study distinct communities, understand the relationship between offline behaviors and online discussion, and clarify social network dynamics that affect physical activity (De Choudhury 2014; De Choudhury, Counts, and Horvitz 2013; De Choudhury et al. 2013; De Choudhury, Sharma, and Kiciman 2016; Dos Reis and Culotta 2015; Eichstaedt et al. 2015; He et al. 2013; Park et al. 2016; Turner-McGrievy et al. 2013).

Social media studies show that language selection appears related to health outcomes. Twitter is a potentially powerful data source, but requires attention to a wide range of issues not often faced with traditional methods. Eichstaedt et al. (2015) and Dos Reis and Culotta (2015) find that using positive language in tweets was a protective factor for health outcomes including heart disease and depression. Also, Gore, Diallo, and Padilla (2015) discovered geographic differences in obesity rates based on the overall discussion and physical activity tweet intensity. In sum, these social media studies assessing physical activity differences reveal that language is an important behavioral predictor and online behavior is related to offline behavior. Twitter is a powerful medium with the requisite diversity and scale to clarify associations between demographic background and physical activity attitudes.

2.2.7 Hypotheses

Multiple hypotheses emerge from the literature on physical activity attitudes and social media data in demographic research. Literature reviewed on racial, gender and intersectional variation in physical activity attitudes suggests the following attitudinal differences:

1. Men will exhibit more positive physical attitudes than women

- Hayes et al. (1999); Eyler et al. (1998)

2. Racial minorities (Black, Asian, Hispanic, and "Other") will report less positive attitudes towards physical activity than Whites

- Baptiste-Roberts et al. (2007); Egede and Bonadonna (2003)

3. Racial minority women will have the least positive attitudes of demographic subgroups

- Airhihenbuwa et al. (1995)

2.3 Data and Methods

The data for this study were gathered with Twitter's free Streaming Application Programming Interface (API) from August 2016 to January 2017. Twitter is a social networking service that allows users to message each other globally and/or directly in short microblogging posts known as tweets. Tweets are constrained to 140 characters at a maximum and allow users to document, share, and interact with public and private communities. Twitter's free Streaming API was used to gather tweets because it provides a real-time continuous connection to Twitter and updates on tweets matching search criteria. The Streaming API represents 1% of all tweets and analyses show that this 1% sample is a random representative sample of all tweets (Sloan et al. 2013). Searching for tweets by subject with the Streaming API means that tweets are not filtered by location (conversely, filtering for tweets for location precludes searching for tweets by hashtag). Using Amazon Web Services (AWS) Elastic Computing (EC2) server, I collected English-language tweets

almost daily for essentially the entire day[5]. Tweets were gathered, analyzed, and processed with R software (R Core Team 2016); this analysis relied heavily on the `UScensus2010`, `ggmap`, `streamR`, `twitteR`, `wru`, `stringr`, and `dplyr` packages (Almquist 2010; Barbera 2014; Gentry 2015; Kahle and Wickham 2013; Khanna and Imai 2016; Wickham 2016; Wickham et al. 2017)

2.3.1 Physical Activity Attitude Measurement

Initial search terms for physical activity tweets used standards for moderate to vigorous (henceforth, MVPA) created by Godin and Shephard (1985). MVPA are ideal for investigation because these health behaviors are more universally recognized, data generated is less context dependent, and activities are more race-neutral[6]. Additionally, MVPA have near universal recognition through common activities such as walking and biking. The search terms continuously queried on Twitter's Streaming API first included the following specific activities: #biking, #jogging, #pullups, #pushups[7], #running, and #walking. Words or phrases that are preceded by the hashtag symbol (#) create a searchable link to other users that are describing their experience similarly. This shared experience is integral to Twitter and searching for physical activity tweets with the hashtag sharply differen-

---

[5]Tweets were downloaded from August 25th 2016 to January 31st 2017. Tweets were not collected during this period when the server was interrupted for maintenance or the stream to Twitter's API timed out unexpectedly.

[6]Data generated describing physical activity are more likely to include context clues like duration while being less inherently context dependent than other health behaviors that impact chronic disease prevalence. For instance, nutrition consumption is much broader in scope and difficult to investigate without some contextual knowledge (e.g., food proportions or servings). Lastly, health behaviors related to physical activity may be more race-neutral than other health behaviors. Guthman (2008) discusses how race affects the alternative food provision market and produces minority exclusion because "these spaces tend to hail white subjects, whites continue to define the rhetoric, spaces, and broader projects" (395). Cultural and socioeconomic boundaries in nutrition discourse suggest that studies into lay nutrition discussion will be segmented by race and class (Lamont and Molnár 2002).

[7]#pushups was ultimately removed from the analysis for reasons discussed in the **Challenges** section.

tiates users that tried to create a social dialogue about their physical activity instead of incidentally mentioning physical activity keywords (e.g., "running late to work"). Additionally, because individuals can pursue physical activity in facility-based or home-based settings, the following terms were also added to capture home-based physical activities: #homeworkouts, #bodyweightworkouts, #bodyweightexercises (Foster et al. 2005).

2.3.2 Predicting Demographic Background

After collecting almost 830,000 tweets from approximately 230,000 users[8] with the relevant search terms, facial recognition software and a Census-based last name by location predictor were used to estimate Twitter users' demographic background. (Faceplusplus.com (henceforth Face++) generated demographic estimates of race, gender, and age. Face++ is a computer vision software platform that uses an image to predict age (continuously; with a range) as well as gender and single-race (both categorical; with numeric confidence estimates)[9]. Bakhshi, Shamma, and Gilbert (2014) and Rhue and Clark (2016) examined the correspondence between Face++ demographic estimates and humans and found greater than 90% agreement between automatic classifications from Face++ software and human classifications from Amazon Mechanical Turk (MTurk)[10]. Additionally, Rhue and Clark (2016) found the confidence level from Face++ estimates mattered as lower confidence estimates were more likely to have disagreement with human classifications suggesting using thresholds with Face++ data instead of all estimates produced by

[8]193,515 tweets from 3,249 presumably professional accounts were removed. Users were determined to be professional accounts by examining user screen names and users with organizations in their user name (e.g., "News", "Watch Reviews", "Bowling Club", etc.) were removed.

[9]Face++ is moving all operations completely to new API on May 1st 2017 and the current version of the new API does not estimate age range or race nor discuss the details of the new machine learning algorithm approach to detect demographics. This study uses the old Face++ API and the following link describes demographic results produced from that method: `http://old.faceplusplus.com/detection_detect/`.

[10]Researchers are leveraging MTurk, an online marketplace that matches task requesters (researchers) and task completers (subjects), to collect inexpensive, high-quality data (Buhrmester, Kwang, and Gosling 2011).

the software. Face++ demographic background estimates have also fared well in social media based studies by Huang, Weber, and Vieweg (2014), Yadav et al. (2014), and Jang et al. (2016). This study also applied a predictor developed by Imai and Khanna (2016) to predict Twitter users' racial backgrounds. These predictions use the Twitter account last name (based on account screen name) and cross-referencing Census data to estimate the user's racial background. A multidimensional view of Twitter users' race is created by supplementing the racial category image Face++ estimates with the surname estimates when available[11]. 73.47% of the race observations had the same estimate with `wru` package and Face++ (this number increases to 78.04% when considering Face++ does not estimate Hispanicity or 'other' race). When the `wru` and Face++ estimates disagreed, the racial background defaulted to the `wru` racial background estimate to further take advantage of the location data.

Twitter users' demographic background was estimated by sending users' profile picture URL to the Face++ API resulting in estimates for 147,178 tweets from 53,910 users. Exploratory analyses suggest that an individual's demographic estimate from Face++ was impacted by the presence of multiple individuals in a profile and Black and Asian users age seemed underestimated. Software limitations suggest that this analysis is less likely to include individuals (perhaps, those with families) more likely to have group photos. The final analysis was subset to only include tweets with non-missing estimates for relevant demographic characteristics and other exclusionary criteria, most importantly, a location-based restriction. These exclusionary criteria included users with estimated age greater than 18 (20,202 observations dropped). The age range for respondents in this study spans ages 18 to 66. The second exclusion applied was the Face ++ race confidence estimate

[11]The Imai and Khanna (2016) method is a Bayesian predictor that provides racial background probabilities given last name, age, gender, and geographic location such as county, tract group, or block group. This analysis used the last name, age, gender, and county-level predictions.

greater than 50% confidence[12] (3,533 observations dropped). The final exclusionary applied is locational, only tweets with geo-located addresses converted from the latitude and longitude of the user's registered profile. Less than 1% of users have their location data available in coordinate form. Applying the location restriction created a final analytical sample of 4,527 tweets from 1,201 users.

2.3.3 Face++ Examples

Figure 2.1 demonstrates the Face++ API estimating the race, gender, and age of W.E.B. DuBois in 1918 (age 50) and Abraham Lincoln in 1863 (age 54)[13] with the two versions of Face++ API. With the legacy API (version 2), Face++ estimates that DuBois is Black (98.63% confident; DuBois did identify with being Black), male (100% confident), and age 38 (with a range of 10 years)[14]. While the software is accurate (and confident) with the race and gender estimates, the age estimate does not include DuBois' true age (50), although his true age is near the max of the suggested age range (48).

The current Face++ API (version 3) estimates that Lincoln is 67 year-old, White male (Lincoln did self-identify as a white male). The current API does not include a range for the age, race, or gender variable like the legacy version. This API also significantly overestimates Lincoln's age[15].

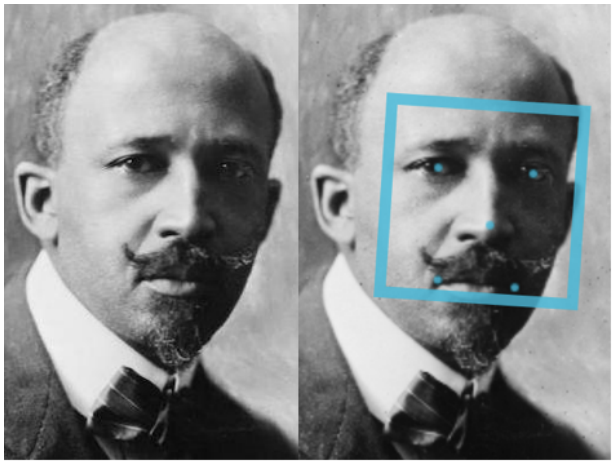
---

[12]The minimum gender confidence estimate was limited to 50% by probability.

[13]The photos used in this example are a photo of W.E.B. DuBois (aged 50) and Abraham Lincoln (aged 54) under Creative Commons license from Wikipedia: `https://upload.wikimedia.org/wikipedia/commons/1/12/WEB_DuBois_1918.jpg` and `https://upload.wikimedia.org/wikipedia/commons/a/ab/Abraham_Lincoln_O-77_matte_collodion_print.jpg`.

[14]Go to the API Demos at `http://old.faceplusplus.com/demo-detect/` and `https://www.faceplusplus.com/attributes/#demo` and use the image link from the previous footnote to see Face++ estimates. The old API is not guaranteed to work after May 1st 2017.

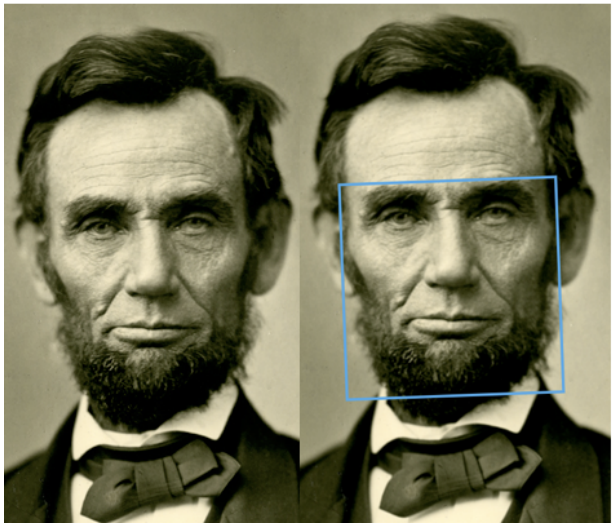
[15]A comparison of the same Lincoln image with the legacy API shows suggests that Lincoln's age was 50 with a range of 10 (gender and race had confidence estimates were greater than 99.7%). Similarly, in the new API, DuBois is estimated to be a 45 year-old Black male. Although a systematic review has not been completed, the current version (version 3) of the Face++ API appears to overestimate age while the



W.E.B. DuBois
1918, age 50

Face++ API
Version 2
Facial Attribute Detection
Demographic Estimates Highlighted

width	28.04%
height	21.50%
center	(61.85%, 27.08%)
eye_left	(55.06%, 22.95%)
eye_right	(69.70%, 23.68%)
nose	(63.03%, 29.74%)
mouth_left	(55.00%, 36.21%)
mouth_right	(66.85%, 37.10%)
age	38 (±10)
race	Black (98.63%)
gender	Male (100.00%)
glass	None (97.94%)
smiling	6.63%
roll_angle	3.73°
yaw_angle	22.02°
pitch_angle	-0.00°



Abraham Lincoln
1863, age 54

Face++ API
Version 3
Facial Attribute Detection
Demographic Estimates Highlighted

Age
67
Gender
Male
Smiling
value: 1.28; threshold: 30.1
Head Pose
Emotion
neutral
Ethnicity
White

Figure 2.1: Face++ API Examples

2.3.4 Analytical Strategy

Main Analysis

This study analyzes physical activity attitudes expressed in tweets with sentiment analysis. Sentiment analysis uses computationally intensive techniques to identify positive, neutral or negative opinions in text (Pak and Paroubek 2010). Computational approaches to opinion, sentiment, and subjectivity in text emerged from natural language processing, a computer science field at the intersection of computation and linguistics (Agarwal et al. 2011). Natural language processing is concerned with using computers to understand human text and speech (Chowdhury 2005).

Multiple textual features are used to investigate tweets sentiment including lexical features, part-of-speech features, n-gram features and micro-blogging features. Lexical features are concerned with sentence level polarity (positive, negative, neutral), part-of speech features include number of verbs, adverbs, adjectives, nouns, and any other parts of speech. Furthermore, n-gram features are a contiguous sequence of n items from a tweet and micro-blogging features capture the presence of positive, negative, and neutral emoticons and abbreviations and the presence of intensifiers (e.g., all-caps and character repetitions). While there are strategic advantages and disadvantages to each textual feature, prior research has established lexical features are a good representation of Twitter sentiment, especially in comparison to other metrics such as part-of-speech features (Kouloumpis, Wilson, and Moore 2011). For an in-depth review of the sentiment analysis and its origins in natural language processing see Pang and Lee (2008).

Sentiment analysis use an opinion lexicon and scoring algorithm to assign a single numeric score to a body of text (here tweets). The opinion lexicon consists of select positive and negative words with predefined scores. The scoring algorithm produces a single sentiment score for a body of text by subtracting the values corresponding to negative words from positive words (positive words score +1, negative words score -1)

---

previous version (version 2) tended to underestimate age.

found in the opinion lexicon[16]. This study uses an opinion lexicon with nearly 6,800 words created by Hu and Liu (2004) and Liu, Hu, and Cheng (2005) and implements the Breen (2012) scoring algorithm. Negative and positive scores correspond to negative and positive opinions; respectively. In this study, the sentiment analysis indicate how positive or negative individuals feel about various physical activities. Example tweets were sampled from the minimum and maximum sentiment score of each racial group and then selected to preserve anonymity by favoring retweets or broad, general language. The example tweets are shown in the **Findings** section.

Synthetic Data

This study leverages synthetic data to meet improve reproducibility and acknowledge contractual agreements. Synthetic data simultaneously allows for code and results to be shared in replicable fashion[17] and also comply with the Twitter terms and conditions for redistribution. Furthermore, by introduce additional noise to social media data, the synthetic data is further anonymized from the actual users granting Twitter users a layer of privacy. The synthetic data sets were created by using the open source software DataSynthesizer. DataSynthesizer can synthesized data in three different modes with that add varying degrees of differential privacy (noise) (Ping, Stoyanovich, and Howe 2017). The basic approach for each mode is the same (create a distribution for each variable and samples that distribution); however, the degree of differential privacy and missing rates used to further anonymize the data change. The software for creating synthetic data has a noise parameter that is equivalent to the minimum change in the correlations by removing (perturbing) any tuple of data at random. Missingness is determined by the synthetic mode utilized (See Appendix A.2 for a description of the synthetic data modes) The independent attribute mode of data synthesis was used to create the synthetic data. All

[16]Tweets also can include emoticons (emojis) and researchers are developing methods to determine sentiment expressed by these images (Kiritchenko, Zhu, and Mohammad 2014). This study does not analyze emojis.

[17]All code and data will be available at `https://github.com/kpolimis`.

sentiment and polarity analyses by demographic groups use synthetic data while analyses with parts of text (e.g., discussion of tweet text or hashtag-specific subjectivity scores) use actual Twitter data (not released).

Sensitivity Analysis

The literature reviewed shows that studies using social media should be aware of potential influences from: managed presentation of self, user selectivity, geographic filters, subject reliability, and demographic reliability. Sensitivity analyses explored potential biases from users "curating" digital traces as well as software and computational limitations from Twitter and supporting data sources.

2.4 Findings

2.4.1 Summary

Similar to previous studies that use Twitter data, I find that my population is more demographically homogeneous than the United States population-at-large. The summary statistics table (Table 2.1) show that users in my dataset are relatively young adults (average age = 35.34) and the sample is less racially-diverse (82% White, 9 Asian%, 6% Hispanic, 2% Black and 2% Other). Additionally, men are over-represented in this sample (55% of the sample is male). The search terms tracked (e.g., physical activity may skew young) could also influence the demographic homogeneity displayed in the sample. See Figure 2.2: "Race by Gender Distribution of Physical Activity Hashtags" for hashtag counts by race and gender (Appendix A.3 includes tabular representation of the same data).

Polarity Scores

Before exploring the findings, a quick note on sentiment analyses: the focus on sentiment analysis is the broad contours of the opinion captured with measures such as polarity and inequality. However, research is unsure which aggregate measures of difference are best (O'Connor et al. 2010). Sentiment analyses will evolve as they link sentiment with a behavioral ground-truth (e.g., steps walked). For instance, Althoff et al. (2017) show that inequality in steps is a better predictor of obesity than average activity volume demon-

strating how the focus is on the variation. Lastly, Gonçalves et al. (2013), show that while sentiment methods perform better with expression-laden content such as social media data (e.g., Twitter), sentiment methods found social media more positive than other text in a way that could reflect Twitter business practices like limiting negative sentiment tweets from discovery with the streaming API. This study uses the definition of polarity and subjectivity established by Godbole, Srinivasaiah, and Skiena (2007):

$$\text{Polarity}_{rg} = \frac{\text{positive references}_{rg}}{\text{total references}_{rg}}$$

$$\text{Subjectivity}_{rgh} = \frac{\text{hashtag references}_{rgh}}{\text{total references}_{rg}}$$

$$\text{for r=race, g=gender, h=hashtag}$$

The findings from the sentiment analysis indicate demographic variation in health behavior attitudes inconsistent with the all demographic-based hypotheses. However, t-tests of sentiment scores did not reveal statistically significant different scores across racial groups or in the intersectional analysis and the discussion is centered on the magnitude of differences between polarity scores. For instance, prior demographic research suggests that men express more positive attitudes towards physical activity than women (Hypothesis 1). Nearly half (47.12%) of the data are tweets with a sentiment score of 0. Analyses revealed that women (0.21) and men had identical polarity scores (0.2). This finding did not support the first hypothesis regarding gender variation towards physical activity attitudes.

Furthermore, the intersectional race and gender comparisons revealed important racial differences in attitudes towards physical activity[18]. Research hypothesizes that

[18]Aging research hypothesizes that similar barriers to exercise and increased awareness of exercise benefits may minimize attitudinal variation in older adults relative to younger adults (King 2001; Mathews et al. 2010; Motalebi et al. 2014). This study analyzed demographic group sentiment by United Nations age categories (e.g., 15 to 19, 20 to 24, etc.) and found that older age groups reported the least positive physical

Whites will have more positive physical activity attitudes than racial minorities (Hypothesis 2). Whites (0.2) and Hispanics (0.2) had the lowest polarity scores, with Asians, Blacks, and Other race reporting the same overall polarity (0.24). These demographic findings are largely inconsistent with the second demographic hypothesis (See Table 2.6 for hypotheses review). Similar to the main analysis, intersectional analyses revealed that White and Hispanic women had the lowest (positive) polarity while Black (0.26) and Other race women (0.3) reported the highest polarity (Table 2.1). Minority women did not have the lowest polarity of all demographic sub-groups refuting the third hypothesis. Additional race by gender analyses revealed that Other race males report the lowest polarity (0.16), in ascending polarity, White males (0.2), Hispanic males (0.21), Black males (0.22), and Asian males (0.24).

Table 2.1: Intersectional Analysis of Polarity Scores

Gender	Race	Total	Proportion	Mean Age	Polarity
Male	White	2017	0.45	35.25	0.2
Female	White	1676	0.37	35.44	0.2
Male	Asian	209	0.05	35.42	0.24
Female	Asian	184	0.04	34.61	0.24
Male	Hispanic	175	0.04	34.3	0.21
Female	Hispanic	111	0.02	36.1	0.18
Male	Black	49	0.01	32.98	0.22
Female	Black	38	0.01	33.71	0.26
Female	Other	37	0.01	34.27	0.3
Male	Other	31	0.01	37.06	0.16

attitudes and lower variation in these attitudes. Individuals aged 44-49 were the most positive age group while individuals aged 35-49 tended to be more positive than all other age groups (See Appendix A.5 for full age analysis results).

Textual investigations beyond sentiment and polarity scores demonstrated the contrast between some tweets that were physical activity related and those that used physical activity keywords but were not related to physical activity. The following positive and negative tweets are emblematic of many tweets in the data set, that is, tweets with physical activity keywords and topically discussing physical activity. (See Appendix A.2; tweets were selected by subsetting the maximum and minimum sentiment scores and filtering in special instances that are discussed in the **Challenges** section)

2.4.2 Example tweets

Black men:

> **Positive**: "13 miles through the rain. Happy I got it done! #run #running #marathontraining"

White women:

> **Positive**: "My happy place! Thankful to be back doing what I love. #RUNNING #fitness #gym #love"

Asian women:

> **Negative**: "Since my knee injury I refuse to to anything that includes #running or walking long distance!"

2.4.3 Average Polarity by Individual Physical Activity

Individual activity analysis reveals variation between demographic groups and physical activities. For instance, racial groups displayed more positive polarity on some physical activities than others that may influence the results. The subjectivity scores show that running-related tweets make up 81.58% of all tweets (98% of all tweets mention #running, #walking, #jogging or #biking) and show key racial and gender differences (See Appendix A.3 for activity counts/proportions disaggregated by demographic group and Appendix A.4 for all activity-specific subjectivity scores disaggregated by demographic group). On average, non-running-related tweets are discussed more positively than running tweets

(Table 2.2). Women have more positive polarity with non-running activities than men and Black females write more positively about running than all other demographic subgroups. In the running-only tweets, all groups with the exception of Black women have a low polarity. Black women have the highest polarity overall and this trend holds for running-only and non-running tweets. The relationship between running and Black female tweets help drive the overall findings that Black females have the most positive attitudes towards physical activity. Running tweets have the greatest subjectivity for all racial groups (proportional relevance with other topics) and black women have by far the greatest polarity for running-tweets. Further examination of tweet sub-samples[19] suggested that the discourse around running is negative because of external considerations (e.g., cold weather, difficult terrain, physical safety concerns, thought about worries while running, etc., See Appendix A.2 for race by gender tweets that discuss these themes) suggesting that attitudes towards running are not reflective of the actual physical activity.

Table 2.2: Running and Non-Running Polarity Scores

Gender	Race	Running Polarity	Non-running Polarity
Female	Asian	0.02	0.23
Female	Black	0.16	0.21
Female	Hispanic	0.05	0.12
Female	White	0.08	0.12
Male	Asian	0.03	0.02
Male	Black	0.01	0.03
Male	Hispanic	0.06	0.25
Male	Other	0	0

[19]Sub-samples were created by randomly sampling 100 to 1000 tweets within demographic groups. The running versus non-running analysis leverages the actual Twitter data so the overall sentiment produced by summing these analyses will not resemble the summary statistics created from synthetic data in Table 2.1.

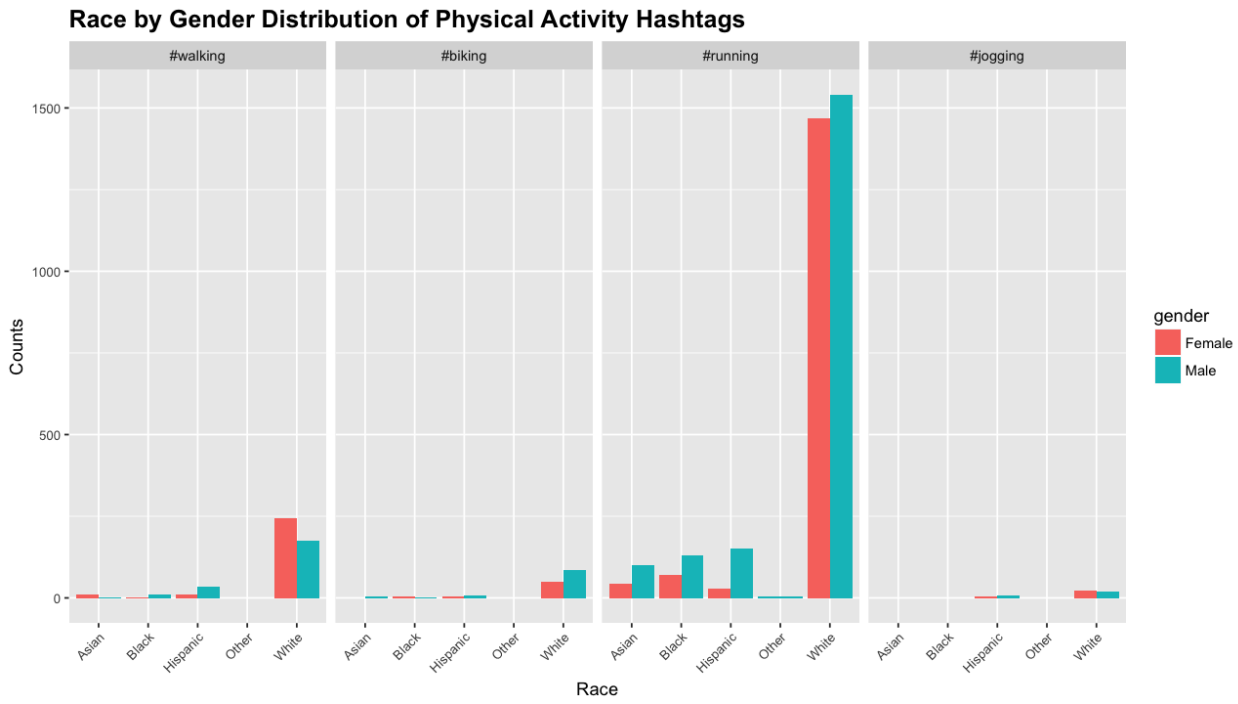


Figure 2.2: Race by Gender Hashtag Histogram

2.4.4 Sensitivity Analysis

Presentation of self

Retweets[20] were more positive than mentions and original tweets for all racial-gender combinations except Black females (original tweets greater polarity) and Black males (tweets with mentions had highest polarity). Mentions had more positive polarity than original tweets for Black females, Asian females, Hispanic males, and Black males. Only original tweets from Black females had greater polarity than other types of tweeting. These findings suggest that when Twitter users apply the words of other individuals to describe their attitudes towards physical activity users will express increased positive

[20]retweets: sharing someone else's tweet; tweet: sharing one's own thoughts to all Twitter users; mentions: sharing one's thought with a specific user in mind (can also be visible to other Twitter users) from `https://support.twitter.com/articles/166337`.

polarity towards physical activity than when these same individuals author tweets to specific individuals or tweet at no one in particular (All sensitivity analysis tables are available in Appendix A.5).

User selectivity

Tweets created by individuals with only tweet in the analysis were more positive towards physical activity for all female subgroups except Hispanic females. Conversely, all male subgroups that had multiple tweets were more positive in their polarity towards physical activity than males with only one tweet. The strong gendered pattern can be a combination of Twitter user behavioral selectivity and actual differences in attitude that are worth exploring in future research that considers physical activity frequency.

Demographic reliability

The sensitivity of Face++ racial and gender confidence estimates were also explored. Comparisons were made at the 99th, 95th, 90th, 85th, 80th, and 50th percentile to understand how increasing racial and gender accuracy affected results. These analyses showed remarkable consistency in polarity scores for demographic subgroups from the 50th to 99th percentile; the 50th percentile was marginally more positive for Black females and Asian males while remaining virtually unchanged for all other racial groups (Table 2.3). The gender confidence analysis showed more variation than the racial confidence interval, but maintained the trend of striking similarity between polarity scores at the upper- and lower-bound of gender confidence estimates. The gender reliability analysis indicated marginally more positive polarity at the 50th than 99th percentile for all racial subgroups except Hispanic males and females. The finding that Hispanic males and females express more positive polarity at the 99th than 50th percentile suggests digital methods to improve identification of Hispanicity will be important for determining demographic differences. Lastly, at all percentiles of gender confidence, the results from the main analysis were consistent for demographic groups (e.g., White females were the most positive subgroup from 99th to 50th percentile, Asian females and males alternate having least positive attitudes across percentiles for any subgroup).

Table 2.3: Polarity by Racial Confidence

Gender	Race	99%thPercentile Polarity	50%thPercentile Polarity
Female	White	0.09	0.09
Female	Black	0.16	0.17
Female	Asian	0	0.07
Female	Hispanic	0.1	0.08
Male	White	0.1	0.09
Male	Hispanic	0.33	0.12
Male	Black	0.01	0.02
Male	Asian	0	0.03

Subject reliability

Using a new set of physical activity subjects (#mma, #boxing, #basketball, #crossfit, #workout, #weightlifting, #wrestling, #golf, #tennis, #skiing, #horsebackriding, and #yoga), demographic variations in polarity was also assessed. These new hashtags revealed different results from the hashtags in the main analysis (Table 2.4). The following findings from the sensitivity analysis were not observed in the main analysis: men had more positive attitudes than women (Hypothesis 1 supported), racial minorities reported equally positive attitudes with Whites (Hypothesis 2 not conclusive). Lastly Hypothesis 3, racial minority women would the lowest polarity was observed in the data. Changes in subject matter had significant effects on the polarity observed.

Table 2.4: Subject Reliability: Intersectional Analysis of Polarity Scores

Gender	Race	Total	Proportion	Mean Age	Polarity
Female	White	72997	0.48	26.85	0.09
Female	Black	3131	0.02	33.02	0.11
Female	Asian	7716	0.05	26.45	0.11
Male	White	50505	0.33	34.73	0.14
Male	Asian	5686	0.04	31.43	0.12
Male	Black	12450	0.08	33.32	0.1

Geographic filters

Geo-located tweets within the US were more positive than non-US geo-located for only Black females (White males were even in both context). While the non-US tweets are not able to estimate Hispanicity and "other" race, Asian males and females, White females, and Black males all discussed physical activity with more positive polarity outside the US. A second geographic filter was created by leveraging the Zillow House Search API[21] to understand the relationship between socioeconomic status (SES) and attitudes towards physical activity. Home values were grouped into three tiers by determining the home values that corresponds to various income groups (group 1: lower-income [household income less than $42,000], group 2: middle-income [household income greater than $126,000 > $42,000], group 3: high-income [household income greater than $126,000]) and clear SES differences emerged in the two demographic groups that were well represented across SES categories. Table 2.5 shows the demographic subgroups with at least 5

[21] Zillow is a real estate marketplace that provides home valuations through the House Search API `https://www.zillow.com/howto/api/GetSearchResults.htm`. A tweet's geo-location was reverse-geocoded into a street address with the `ggmaps` package (Kahle and Wickham 2013) and the street addresses were sent to the Zillow API for home value estimates.

individuals in each SES category and is mostly about White males and females (although the patterns described apply broadly to all racial groups in these SES categories). The income pattern is most straightforward in White males where individuals with greatest value homes expressed more positive polarity than individuals from low- and middle-income areas. For White women, high income homes also were associated with the most positive polarity, however, physical activity tweets from low-income areas had a greater polarity than tweets from a middle-income area. In total, tweets occurring from users associated with high-income home values were on average more positive than tweets from other income backgrounds. The small sample sizes for estimate household income make it difficult to test the hypothesis by Ray (2017) that intersectionality has differential effects for minorities based on variations in neighborhood experience.

Table 2.5: SES Analysis of Polarity Scores

Gender	Race	SES	Total	Polarity
Female	White	3	96	0.19
Female	White	2	102	0.08
Female	White	1	11	0.18
Male	White	2	149	0.05
Male	White	3	108	0.13
Male	White	1	44	0
Male	Black	2	14	0
Male	Hispanic	3	6	0

Table 2.6: Hypothesis Table

Hypothesis	Finding
1. Men will show more positive attitudes than women	Not supported
2. Racial minorities (Blacks; Asians; Hispanics; and Other race) will report less positive attitudes towards physical activity than Whites	Not supported
3. Female racial minorities will have the least positive attitudes of demographic subgroups	Not supported

Sensitivity approaches illustrate the how Twitter data is influenced by multiple data, audience, and demographic limitations (Table 2.7). In the sensitivity analyses, virtually every sensitivity examined (tweet audience, geographic filters, hashtags, demographic filters, user selectivity) changed the findings from the main analysis in some way. On average, the sensitivity analyses minimized the variation between demographic subgroups. The least significant changes were gender confidence filters while the most significant changes were produced by shifting tweet audience (original tweets vs retweets vs mentions) and applying geographic filters. These findings suggest that establishing thresholds and filters for "usable data" (e.g., racial confidence thresholds, geographic filters) can improve the accuracy of claims.

Table 2.7: Sensitivity Analysis

Approach	Overall Findings	Changes to Race Findings	Changes to Gender Findings
Presentation of self	Retweets[22] more positive than original tweets or mentions	All races more positive (no change in order)	No changes
User selectivity	Individuals with one tweet more positive than users with multiple tweets	No changes	No changes
Subject reliability	Terms that define "physical activity" important	Blacks and Asians more positive than Whites	Men more positive than women
Demographic reliability	No changes from main analysis	No changes	No changes
Gender reliability	No changes from main analysis	No changes	No changes
Geographic filter (US only tweets)	US tweets different from non US tweets	Blacks and Asians more positive than Whites	No changes
Geographic filter (SES proxy)	Higher value homes associated with more positive sentiment	See Appendix E: small sample sizes	See Appendix E: small sample sizes

[22]retweets: sharing someone else's tweet; tweet: sharing one's own thoughts to all Twitter users; mentions: sharing one's thought with a specific user in mind (can also be visible to other Twitter users) from `https://support.twitter.com/articles/166337`.

2.5 Conclusion

Some demographic variations in MVPA attitudes were observed in this analysis. Tweet polarity analysis revealed racial similarities and differences where Whites and Blacks were equally positive in their discussion of physical activity and more positive than Asians. Gender analysis revealed that women had more positive polarity than men towards physical activity; the racial and gender results slightly support previous demographic research although the gender findings are more inconsistent than the racial findings. Whites and Blacks had a near equal gender gap in polarity (.07 and .08) while Asians had the smallest gender gap (.02).

The racial differences in attitudes towards physical activity shown in this study were largely driven by White females and Black males displaying the most positive attitudes towards physical activity. The relatively positive polarity of Black users relative to White males and females complicates the established relationship between attitudes and physical activity engagement. Attitudinal research suggests that positive health attitudes should correlate with more engagement in physical activity and ultimately lower chronic disease burden; however, results from Black users challenge this established attitudinal relationship. Black females positive attitudes towards physical activity juxtaposed with Black males less positive attitudes adds additional complexity to the attitude-engagement relationship because these subgroups have contrasting attitudes toward physical activity but similar physical activity trends. However, these findings regarding physical activity are sensitive to multiple Twitter data limitations.

Sensitivity approaches revealed that Twitter data is subject to audience effects (presentation of self), geographic limitations, and subjects reliability. For instance, while the main analysis shows that Black females have a top two average sentiment and Black females have bottom three average sentiment, geographic comparisons call this finding into question. In US geo-located tweets, Black females are much more positive than Black

males while the inverse is true for the non US geo-located tweets. Future research should strive to incorporate methods that acknowledge the limitations of Twitter that the sensitivity analyses demonstrated while also addressing the potential seasonality (e.g., activity preferences and resulting attitudes towards physical activity could change based on time of year). A wider date range in tweets that included several months could remove the potential biases from five months.

2.6 Challenges

While investigating the tweets, multiple potential sources of spuriousness related to Twitter data were identified that suggest further data pre-processing is necessary before calculating sentiment scores. First, specific hashtag searches on Twitter's Streaming API cannot be bound to specific geographic areas. Secondly, due to Twitter's inherent social atmosphere, social phenomena occurring at the same time can appropriate hashtags from other movements. Two hashtags, #pushups and #walking, are emblematic of the locational limitations and discussion conflation that can occur with Twitter data.

#pushups

On October 31st 2016, Imran Khan,leader of the Pakistan Tehreek-i-Insaaf (PTI) party, gained international (viral) attention[23] for doing pushups as a sign of strength before a planned government protest on November 2nd. Imran Khan's followers wrote scathing posts about the Pakistani government and included #pushups in their messages. Given the inability to search for subjects (by hashtag(#) or keyword) and simultaneously limit geographic area, misidentified pushups tweets complicated this study and ultimately all #pushups tweets were removed from the final analysis (See Appendix A.5 for sentiment analysis of removed #pushups tweets).

---

23 http://timesofindia.indiatimes.com/world/pakistan/Imran-Khan-does-50-push-ups-as-warm-up-for-November-2-Islamabad-protest/articleshow/55153350.cms

#walking

Another source of spurious tweets was related to the social interaction that Twitter tries to foster. For instance, viewers of the popular television show Walking Dead use "#walkingdead" to participate in discussions around the television show. However, some Walking Dead viewers used "#walking dead" which is different from "#walkingdead". Because both hashtags to discuss the television show include "#walking", the subject filter gathered these tweets. Ultimately, both instances for referring to the television show Walking Dead were removed from the data. Removing misidentified tweets and applying further scrutiny to the tweet text improves estimates of demographic differences in physical activity.

Third, Twitter users are not obligated to put up a recent or factual photograph as a profile picture. The anonymity provided by the internet makes it easy to represent oneself in anyway possible such as pretending to be someone else in online social networks. Users with a fake profile picture could easily confound Face++ demographic background estimates. Fourth, professional organizations communicate with followers via Twitter and early analysis showed that some profile pictures on accounts from these organizations were likely fitness models. These tweets potentially confound the sentiment analysis by providing more positive phrases and increase demographic skewness in representation. However, most individuals are likely to include a picture of their likeness and the analysis removed users with professional organizations in their name. The combination of user freedom in profile picture selection and professional tweets are the largest driving force of attrition in the study and responsible for more than 50% of the initially downloaded tweets not entering the final analytical sample. Fifth, this project specified activity terms that were preceded by a hashtag which limits the sample size. While this method was used to exclude incidental physical activity term mentions (e.g., "running late to work"), intentional using physical activity terms without the hashtag symbol leads to substantively

relevant tweets being ignored[24].

Additionally, the specific physical activities followed may reflect cultural preferences and/or environmental constraints that potentially limit the analysis. For instance, Bourdieu (1993) argues that working- and middle-class individuals pursue physical activity in which the body is used to conquer others while the upper-class pursue physical activity for fitness, which helps their professional goals. To that end, some hashtags (#running, #jogging, and #biking) may represent physical activity markers biased toward the upper class. Beyond cultural tastes potential influences, multiple public health studies have found income differences in physical activity such as walking and biking related to built environment that could potentially affect conclusions (Brownson et al. 2001; Ewing et al. 2003; Giles-Corti 2002; Gordon-Larsen et al. 2006; Hoehner et al. 2005). Sensitivity analyses explored the variation in attitudes towards a broader physical activity set[25] and used SES proxies.

Lastly, Twitter is notorious for activity from "socialbot"'(also known as 'bot') accounts, pre-programmed interactive scripts that appear as humans and "cyborgs" either bot-assisted human or human-assisted bot (Chu et al. 2012; Rouse 2013). Through tailored algorithms, bots and cyborgs can become influential by acquiring numerous followers that retweet content; the extent to which bots are on Twitter and other social media is uncertain (Ferrara et al. 2016; Messias et al. 2013). Its entirely possible that tweets from bots are included in this analysis. Gender dynamics regarding bots may also complicate this analysis. While Freitas et al. (2014) found that there is no significant gender difference in popularity acquired by socialbots in the aggregate, gender was influential

[24]Tweets that use a physical activity keyword, but are not contextually related to physical activity such as "#Trump's #lewd and #obscene talk about #woman,making belittling #vulgar #comments,is last straw on back of the #camel.#Running for #POTUS?" are being removed.

[25]New search terms such as #boxing, #basketball, #crossfit, #weightlifting, #golf, and #tennis were not added midstream during data collection to maintain continuity.

for socially connected users posting on the same topic. Additionally, Shafahi, Kempers, and Afsarmanesh (2016) show that content (especially retweeted and duplicated tweets) originating from female twitter bots is shared more frequently than male bots suggesting that retweeted content from profiles with female pictures have a greater bot likelihood. I will identify and remove bot tweets to improve conclusions about attitudinal variation by using the BotOrNot Service developed by Davis et al. (2016). This open source service leverages Twitter users' recent account history to predict the likelihood the user is a bot.

Ongoing improvements to the project include using approaches created by Barbera (2014) and Vicente, Batista, and Carvalho (2016) to further leverage profile data to examine linguistic and demographic concerns. Barbera has produced materials to gather entire timeline data for users. Entire timeline data could clarify the relationship between the Twitter user's language in general and their language when discussing physical activity. For instance, negative tweets about physical activity could illustrate user's negative language in general and not negative attitudes about physical activity per se. Additionally, Vicente et al. (2016) has introduced a text mining approach with an individual's Twitter user name and screen name to aid in gender detection. Lastly, Nielsen (2011) and Baccianella, Esuli, and Sebastiani (2010) have developed microblogging specific sentiment analysis such as greater valence score ranges[26] that may capture social media dynamics more accurately. Incorporating these approaches will clarify sentiment analysis results and demographic background conclusions.

[26]Valence is the positivity, neutrality, or negativity of an opinion. Traditional sentiment analyses scale valence from -1 (negative) to +1 (positive). Nielsen (2011) proposes rating most positive and negative words +2 and -2, respectively. However, extreme negativity can be rated up to -5.

2.6.1 References

Abiteboul, Serge. 1997. "Querying Semi-Structured Data." Greece.

Affuso, Olivia, Tiffany L. Cox, Nefertiti H. Durant, and David B. Allison. 2011. "Attitudes and Beliefs Associated with Leisure-Time Physical Activity Among African American Adults." *Ethnicity & Disease* 21(1):63–67.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. "Sentiment Analysis of Twitter Data." Pp. 30–38 in *Proceedings of the Workshop on Languages in Social Media, LSM '11*. Portland, Oregon: Association for Computational Linguistics. Retrieved (`http://dl.acm.org/citation.cfm?id=2021109.2021114`).

Airhihenbuwa, Collins O., Shiriki Kumanyika, Tanya D. Agurs, and Agatha Lowe. 1995. "Perceptions and Beliefs About Exercise, Rest, and Health Among African-Americans." *American Journal of Health Promotion* 9(6):426–29. Retrieved August 10, 2016 (`http://ajhpcontents.org/doi/abs/10.4278/0890-1171-9.6.426`).

Ajzen, Icek. 1991. "The Theory of Planned Behavior." *Organizational behavior and human decision processes* 50(2):179–211. Retrieved August 10, 2016 (`http://www.sciencedirect.com/science/article/pii/074959789190020T`).

Ajzen, Icek and Christine Timko. 1986. "Correspondence Between Health Attitudes and Behavior." *Basic and Applied Social Psychology* 7(4):259–76. Retrieved August 10, 2016 (`http://www.tandfonline.com/doi/abs/10.1207/s15324834basp0704_2`).

Almquist, Zack W. 2010. "US Census Spatial and Demographic Data in R: The UScensus2000 Suite of Packages." *Journal of Statistical Software* 37(6):1–31. Retrieved (`http://www.jstatsoft.org/v37/i06/`).

Althoff, Tim et al. 2017. "Large-Scale Physical Activity Data Reveal Worldwide Activity Inequality." *Nature* 547(7663):336–39. Retrieved July 24, 2017 (`http://www.nature.`

com/doifinder/10.1038/nature23018).

Arseniev-Koehler, Alina, Hedwig Lee, Tyler McCormick, and Megan A. Moreno. 2016. "#Proana: Pro-Eating Disorder Socialization on Twitter." *Journal of Adolescent Health* 58(6):659–64. Retrieved March 6, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S1054139X16000598).

Azjen, Icek. 1985. "From Intentions to Actions: A Theory of Planned Behavior." in *Action control: From cognition to behavior*, *Springer series in social psychology*, edited by J. Kuhl and J. Beckmann. Berlin: Springer.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." Pp. 2200–2204 in *LREC*, vol. 10.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3-4):465–82. Retrieved August 10, 2017 (http://link.springer.com/10.1007/s11186-014-9216-5).

Bakhshi, Saeideh, David A. Shamma, and Eric Gilbert. 2014. "Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram." Pp. 965–74 in. ACM Press. Retrieved February 17, 2017 (http://dl.acm.org/citation.cfm?doid=2556288.2557403).

Baptiste-Roberts, Kesha et al. 2007. "Family History of Diabetes, Awareness of Risk Factors, and Health Behaviors Among African Americans." *American Journal of Public Health* 97(5):907–12. Retrieved August 10, 2016 (http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2005.077032).

Barbera, Pablo. 2014. *streamR: Access to Twitter Streaming API via R*. Retrieved (https://CRAN.R-project.org/package=streamR).

Bellman, Richard Ernest and Stuart Dreyfus. 1957. *Dynamic Programming*. 1. Prince-

ton Landmarks in Mathematics ed., with a new introduction. Princeton, NJ: Princeton University Press.

Bourdieu, Pierre. 1993. "How Can One Be a Sports Fan." *The cultural studies reader* 2:427–40.

Boyington, J. E. A., D. L. Howard, and D. N. Holmes. 2008. "Self-Rated Health, Activities of Daily Living, and Mobility Limitations Among Black and White Stroke Survivors." *Journal of Aging and Health* 20(8):920–36. Retrieved August 10, 2016 (http://jah.sagepub.com/cgi/doi/10.1177/0898264308324643).

Breen, Jeffrey Oliver. 2012. "Mining Twitter for Airline Consumer Sentiment." *Practical text mining and statistical analysis for non-structured text data applications* 133.

Brenes, Gretchen A., Michael J. Strube, and Martha Storandt. 1998. "An Application of the Theory of Planned Behavior to Exercise Among Older Adults1." *Journal of Applied Social Psychology* 28(24):2274–90. Retrieved August 22, 2016 (http://doi.wiley.com/10.1111/j.1559-1816.1998.tb01371.x).

Brownson, Ross C., Elizabeth A. Baker, Robyn A. Housemann, Laura K. Brennan, and Stephen J. Bacak. 2001. "Environmental and Policy Determinants of Physical Activity in the United States." *American Journal of Public Health* 91(12):1995–2003. Retrieved February 17, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.91.12.1995).

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1):3–5. Retrieved February 21, 2017 (http://journals.sagepub.com/doi/10.1177/1745691610393980).

Bullingham, L. and A. C. Vasconcelos. 2013. "'The Presentation of Self in the Online World': Goffman and the Study of Online Identities." *Journal of Information Sci-*

ence 39(1):101–12. Retrieved November 6, 2016 (http://jis.sagepub.com/cgi/doi/10.1177/0165551512470051).

Cavallo, David N. et al. 2012. "A Social Media–Based Physical Activity Intervention." *American Journal of Preventive Medicine* 43(5):527–32. Retrieved October 31, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S074937971200520X).

Cesare, Nina, Tyler McCormick, Emma S. Spiro, and Emilio Zagheni. 2016. "Promises and Pitfalls of Using Digital Traces for Demographic Research." *SSRN Electronic Journal*. Retrieved October 23, 2016 (http://www.ssrn.com/abstract=2839585).

Chowdhury, Gobinda G. 2005. "Natural Language Processing." *Annual Review of Information Science and Technology* 37(1):51–89. Retrieved November 7, 2016 (http://doi.wiley.com/10.1002/aris.1440370103).

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing* 9(6):811–24. Retrieved February 15, 2017 (http://ieeexplore.ieee.org/document/6280553/).

Clark, Daniel O. and Faryle Nothwehr. 1999. "Exercise Self-Efficacy and Its Correlates Among Socioeconomically Disadvantaged Older Adults." *Health Education & Behavior* 26(4):535–46. Retrieved August 10, 2016 (http://heb.sagepub.com/content/26/4/535.short).

Cossrow, Nicole and Bonita Falkner. 2004. "Race/Ethnic Issues in Obesity and Obesity-Related Comorbidities." *The Journal of Clinical Endocrinology & Metabolism* 89(6):2590–4. Retrieved November 13, 2016 (http://press.endocrine.org/doi/10.1210/jc.2004-0339).

Courneya, Kerry S., Ronald C. Plotnikoff, Stephen B. Hotz, and Nicholas J. Birkett. 2000. "Social Support and the Theory of Planned Behavior in the Exercise Domain." *American Journal of Health Behavior* 24(4):300–308. Retrieved August 24, 2016 (http://openurl.

ingenta.com/content/xref?genre=article&issn=1087-3244&volume=24&issue=4&spage=300).

Courtenay, Will H., Donald R. McCreary, and Joseph R. Merighi. 2002. "Gender and Ethnic Differences in Health Beliefs and Behaviors." *Journal of health psychology* 7(3):219–31. Retrieved August 10, 2016 (http://hpq.sagepub.com/content/7/3/219.short).

Crenshaw, Kimberle. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color." *Stanford Law Review* 43(6):1241. Retrieved October 24, 2016 (http://www.jstor.org/stable/1229039?origin=crossref).

Crespo, Carlos J., Ellen Smit, Ross E. Andersen, Olivia Carter-Pokras, and Barbara E. Ainsworth. 2000. "Race/Ethnicity, Social Class and Their Relation to Physical Inactivity During Leisure Time: Results from the Third National Health and Nutrition Examination Survey, 1988–1994." *American journal of preventive medicine* 18(1):46–53. Retrieved August 10, 2016 (http://www.sciencedirect.com/science/article/pii/S0749379799001051).

Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. "BotOrNot: A System to Evaluate Social Bots." Pp. 273–74 in. ACM Press. Retrieved February 15, 2017 (http://dl.acm.org/citation.cfm?doid=2872518.2889302).

De Choudhury, Munmun. 2014. "Opportunities of Social Media in Health and Well-Being." *XRDS: Crossroads, The ACM Magazine for Students* 21(2):23–27. Retrieved August 10, 2016 (http://dl.acm.org/citation.cfm?doid=2701297.2676570).

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013. "Social Media as a Measurement Tool of Depression in Populations." Pp. 47–56 in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM. Retrieved August 10, 2016 (http://dl.acm.org/citation.cfm?id=2464480).

De Choudhury, Munmun, Michael Gamon, Aaron Hoff, and Asta Roseway. 2013. "'Moon

Phrases': A Social Media Faciliated Tool for Emotional Reflection and Wellness." Pp. 41–44 in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE. Retrieved August 10, 2016 (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6563900).

De Choudhury, Munmun, Sanket Sharma, and Emre Kiciman. 2016. "Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media." Pp. 1155–68 in. ACM Press. Retrieved August 10, 2016 (http://dl.acm.org/citation.cfm?doid=2818048.2819956).

Dietz, W. H. 1998. "Health Consequences of Obesity in Youth: Childhood Predictors of Adult Disease." *Pediatrics* 101(3 Pt 2):518–25.

Dos Reis, Virgile Landeiro and Aron Culotta. 2015. "Using Matched Samples to Estimate the Effects of Exercise on Mental Health from Twitter." Pp. 182–88 in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Retrieved August 10, 2016 (http://www.cs.iit.edu/~culotta/pubs/virgile15using.pdf).

Dwyer, John J. M. et al. 2006. "Adolescent Girls' Perceived Barriers to Participation in Physical Activity." *Adolescence* 41(161):75–89.

Egede, Leonard E. and Ramita J. Bonadonna. 2003. "Diabetes Self-Management in African Americans: An Exploration of the Role of Fatalism." *The Diabetes Educator* 29(1):105–15.

Eichstaedt, Johannes C. et al. 2015. "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality." *Psychological science* 26(2):159–69. Retrieved August 10, 2016 (http://pss.sagepub.com/content/26/2/159.short).

Ewing, Reid, Tom Schmid, Richard Killingsworth, Amy Zlot, and Stephen Raudenbush. 2003. "Relationship Between Urban Sprawl and Physical Activity, Obesity, and Morbidity." *American Journal of Health Promotion* 18(1):47–57. Retrieved February 17, 2017

(http://ajhpcontents.org/doi/abs/10.4278/0890-1171-18.1.47).

Eyler, Amy A. et al. 1998. "Physical Activity and Minority Women: A Qualitative Study." *Health Education & Behavior* 25(5):640–52. Retrieved August 10, 2016 (http://heb.sagepub.com/content/25/5/640.short).

Eyler, Amy E. et al. 2002. "Correlates of Physical Activity Among Women from Diverse Racial/Ethnic Groups." *Journal of Women's Health & Gender-Based Medicine* 11(3):239–53.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. "The Rise of Social Bots." *Communications of the ACM* 59(7):96–104. Retrieved February 15, 2017 (http://dl.acm.org/citation.cfm?doid=2963119.2818717).

Foster, Charles, Melvyn Hillsdon, Margaret Thorogood, Asha Kaur, and Thamindu Wedatilake. 2005. "Interventions for Promoting Physical Activity." in *Cochrane Database of Systematic Reviews*, edited by The Cochrane Collaboration. Chichester, UK: John Wiley & Sons, Ltd. Retrieved October 27, 2016 (http://doi.wiley.com/10.1002/14651858.CD003180.pub2).

Freitas, Carlos A., Fabrício Benevenuto, Saptarshi Ghosh, and Adriano Veloso. 2014. "Reverse Engineering Socialbot Infiltration Strategies in Twitter." *CoRR* abs/1405.4927. Retrieved (http://arxiv.org/abs/1405.4927).

French, David P. et al. 2005. "The Importance of Affective Beliefs and Attitudes in the Theory of Planned Behavior: Predicting Intention to Increase Physical Activity1." *Journal of Applied Social Psychology* 35(9):1824–48. Retrieved September 22, 2016 (http://doi.wiley.com/10.1111/j.1559-1816.2005.tb02197.x).

Friedman, Jerome H. 1997. "On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality." *Data Mining and Knowledge Discovery* 1(1):55–77. Retrieved (http://dx.doi.org/10.

1023/A:1009778005914).

Gentry, Jeff. 2015. *twitteR: R Based Twitter Client*. Retrieved (https://CRAN.R-project.org/package=twitteR).

Giles-Corti, B. 2002. "Socioeconomic Status Differences in Recreational Physical Activity Levels and Real and Perceived Access to a Supportive Physical Environment." *Preventive Medicine* 35(6):601–11. Retrieved February 17, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0091743502911151).

Giles-Corti, Billie and Robert J. Donovan. 2002. "The Relative Influence of Individual, Social and Physical Environment Determinants of Physical Activity." *Social Science & Medicine (1982)* 54(12):1793–1812.

Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. 2007. "Large-Scale Sentiment Analysis for News and Blogs." *ICWSM* 7(21):219–22.

Godin, G. and R. J. Shephard. 1985. "A Simple Method to Assess Exercise Behavior in the Community." *Canadian Journal of Applied Sport Sciences. Journal Canadien Des Sciences Appliquées Au Sport* 10(3):141–46.

Godin, G., P. Valois, R. J. Shephard, and R. Desharnais. 1987. "Prediction of Leisure-Time Exercise Behavior: A Path Analysis (LISREL V) Model." *Journal of Behavioral Medicine* 10(2):145–58.

Goffman, Erving and others. 1978. *The Presentation of Self in Everyday Life*. Harmondsworth London.

Gonçalves, Pollyanna, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. "Comparing and Combining Sentiment Analysis Methods." Pp. 27–38 in. ACM Press. Retrieved July 24, 2017 (http://dl.acm.org/citation.cfm?doid=2512938.2512951).

Gordon-Larsen, Penny, Melissa C. Nelson, Phil Page, and Barry M. Popkin. 2006. "Inequal-

ity in the Built Environment Underlies Key Health Disparities in Physical Activity and Obesity." *Pediatrics* 117(2):417–24.

Gore, Ross Joseph, Saikou Diallo, and Jose Padilla. 2015. "You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content" edited by D. Meyre. *PLOS ONE* 10(9):e0133505. Retrieved August 10, 2016 (http://dx.plos.org/10.1371/journal.pone.0133505).

Grieser, M. et al. 2006. "Physical Activity Attitudes, Preferences, and Practices in African American, Hispanic, and Caucasian Girls." *Health Education & Behavior* 33(1):40–51. Retrieved August 10, 2016 (http://heb.sagepub.com/cgi/doi/10.1177/1090198105282416).

Guthman, Julie. 2008. "'If They Only Knew': Color Blindness and Universalism in California Alternative Food Institutions." *The Professional Geographer* 60(3):387–97. Retrieved August 10, 2016 (http://www.tandfonline.com/doi/abs/10.1080/00330120802013679).

Hagger, Martin S., Nikos L. D. Chatzisarantis, Trudi Culverhouse, and Stuart J. H. Biddle. 2003. "The Processes by Which Perceived Autonomy Support in Physical Education Promotes Leisure-Time Physical Activity Intentions and Behavior: A Trans-Contextual Model." *Journal of Educational Psychology* 95(4):784–95. Retrieved August 10, 2016 (http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.95.4.784).

Hagger, Martin S., Nikos L.D. Chatzisarantis, and Stuart J.H. Biddle. 2002. "A Meta-Analytic Review of the Theories of Reasoned Action and Planned Behavior in Physical Activity: Predictive Validity and the Contribution of Additional Variables." *Journal of Sport and Exercise Psychology* 24(1):3–32. Retrieved August 23, 2016 (http://journals.humankinetics.com/doi/10.1123/jsep.24.1.3).

Harden, A. 2004. "Applying Systematic Review Methods to Studies of People's Views: An Example from Public Health Research." *Journal of Epidemiology & Community Health* 58(9):794–800. Retrieved August 10, 2016 (http://jech.bmj.com/cgi/doi/10.1136/

jech.2003.014829).

Haskell, William L. et al. 2007. "Physical Activity and Public Health: Updated Recommendation for Adults from the American College of Sports Medicine and the American Heart Association." *Medicine & Science in Sports & Exercise* 39(8):1423–34. Retrieved August 23, 2016 (http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005768-200708000-00027).

Hawn, C. 2009. "Take Two Aspirin and Tweet Me in the Morning: How Twitter, Facebook, and Other Social Media Are Reshaping Health Care." *Health Affairs* 28(2):361–68. Retrieved August 10, 2016 (http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.28.2.361).

Hayes, Sean, Peter Crocker, and Kent Kowalski. 1999. "Gender Differences in Physical Self-Perceptions, Global Self-Esteem and Physical Activity: Evaluation of the Physical Self-Perception Profile Model." *Journal of Sport Behavior* 22(1):1–14.

He, Qian, Emmanuel Agu, Diane Strong, Bengisu Tulu, and Peder Pedersen. 2013. "Characterizing the Performance and Behaviors of Runners Using Twitter." Pp. 406–14 in. IEEE. Retrieved August 10, 2016 (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6680503).

Hendelman, D., K. Miller, C. Baggett, E. Debold, and P. Freedson. 2000. "Validity of Accelerometry for the Assessment of Moderate Intensity Physical Activity in the Field:" *Medicine & Science in Sports & Exercise* 32(Supplement):S442–S449. Retrieved August 22, 2016 (http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005768-200009001-00002).

Henderson, Karla A. and Barbara E. Ainsworth. 2003. "A Synthesis of Perceptions About Physical Activity Among Older African American and American Indian Women." *Amer-*

ican Journal of Public Health 93(2):313–17.

Hoehner, Christine M., Laura K. Brennan Ramirez, Michael B. Elliott, Susan L. Handy, and Ross C. Brownson. 2005. "Perceived and Objective Environmental Measures and Physical Activity Among Urban Adults." *American Journal of Preventive Medicine* 28(2):105–16. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0749379704003034).

Hogan, B. 2010. "The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online." *Bulletin of Science, Technology & Society* 30(6):377–86. Retrieved November 6, 2016 (http://bst.sagepub.com/cgi/doi/10.1177/0270467610385893).

Hu, Minqing and Bing Liu. 2004. "Mining and Summarizing Customer Reviews." P. 168 in. ACM Press. Retrieved February 21, 2017 (http://portal.acm.org/citation.cfm?doid=1014052.1014073).

Huang, Wenyi, Ingmar Weber, and Sarah Vieweg. 2014. "Inferring Nationalities of Twitter Users and Studying Inter-National Linking." Pp. 237–42 in. ACM Press. Retrieved February 17, 2017 (http://dl.acm.org/citation.cfm?doid=2631775.2631825).

Im, Eun-Ok et al. 2013. "Racial/Ethnic Differences in Midlife Women's Attitudes Toward Physical Activity." *Journal of Midwifery & Women's Health* 58(4):440–50. Retrieved August 21, 2016 (http://doi.wiley.com/10.1111/j.1542-2011.2012.00259.x).

Im, Eun-Ok et al. 2010. "'A Waste of Time': Hispanic Women's Attitudes Toward Physical Activity." *Women & Health* 50(6):563–79. Retrieved August 22, 2016 (http://www.tandfonline.com/doi/abs/10.1080/03630242.2010.510387).

Im et al. 2012. "A National Internet Survey on Midlife Women's Attitudes Toward Physical Activity." *Nursing Research* 61(5):342–52. Retrieved August 21, 2016 (http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006199-201209000-00005).

Im, Wonshik Chee, Hyun-Ju Lim, Yi Liu, and Hee Kyung Kim. 2008. "Midlife Women's Attitudes Toward Physical Activity." *Journal of Obstetric, Gynecologic & Neonatal*

Nursing 37(2):203–13. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0884217515300630).

Im, O. K. Ham, E. Chee, and W. Chee. 2015. "Physical Activity and Depressive Symptoms in Four Ethnic Groups of Midlife Women." *Western Journal of Nursing Research* 37(6):746–66. Retrieved August 21, 2016 (http://wjn.sagepub.com/cgi/doi/10.1177/0193945914537123).

Im et al. 2012. "'Physical Activity as a Luxury': African American Women's Attitudes Toward Physical Activity." *Western Journal of Nursing Research* 34(3):317–39. Retrieved August 21, 2016 (http://wjn.sagepub.com/cgi/doi/10.1177/0193945911400637).

Imai, Kosuke and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24(02):263–72. Retrieved February 17, 2017 (https://www.cambridge.org/core/product/identifier/S1047198700010962/type/journal_article).

Jang, Jin Yea, Kyungsik Han, Dongwon Lee, Haiyan Jia, and Patrick C. Shih. 2016. "Teens Engage More with Fewer Photos: Temporal and Comparative Analysis on Behaviors in Instagram." Pp. 71–81 in. ACM Press. Retrieved February 17, 2017 (http://dl.acm.org/citation.cfm?doid=2914586.2914602).

Johnson, Mark R. D. 2000. "Perceptions of Barriers to Healthy Physical Activity Among Asian Communities." *Sport, Education and Society* 5(1):51–70. Retrieved November 8, 2016 (http://www.tandfonline.com/doi/abs/10.1080/135733200114433).

Kahle, David and Hadley Wickham. 2013. "Ggmap: Spatial Visualization with Ggplot2." *The R Journal* 5(1):144–61. Retrieved (http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf).

Kandula, Namratha R. and Diane S. Lauderdale. 2005. "Leisure Time, Non-Leisure Time, and Occupational Physical Activity in Asian Americans." *Annals of Epidemiology* 15(4):257–

65. Retrieved November 8, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S1047279704002443).

Kaplan, Andreas M. and Michael Haenlein. 2010. "Users of the World, Unite! The Challenges and Opportunities of Social Media." *Business Horizons* 53(1):59–68. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0007681309001232).

Khanna, Kabir and Kosuke Imai. 2016. *Wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation*. Retrieved (https://CRAN.R-project.org/package=wru).

King, A. C. 2001. "Interventions to Promote Physical Activity by Older Adults." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 56(Supplement 2):36–46. Retrieved August 10, 2016 (http://biomedgerontology.oxfordjournals.org/cgi/doi/10.1093/gerona/56.suppl_2.36).

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. 2014. "Sentiment Analysis of Short Informal Texts." *J. Artif. Int. Res.* 50(1):723–62. Retrieved (http://dl.acm.org/citation.cfm?id=2693068.2693087).

Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. 2011. "Twitter Sentiment Analysis: The Good the Bad and the Omg!" Pp. 538–41 in *Roceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Krause, N. M. and G. M. Jay. 1994. "What Do Global Self-Rated Health Items Measure?" *Medical Care* 32(9):930–42.

Krämer, Nicole C. and Stephan Winter. 2008. "Impression Management 2.0: The Relationship of Self-Esteem, Extraversion, Self-Efficacy, and Self-Presentation Within Social Networking Sites." *Journal of Media Psychology* 20(3):106–16. Retrieved November 6,

2016 (http://econtent.hogrefe.com/doi/abs/10.1027/1864-1105.20.3.106).

Lamont, Michèle and Virág Molnár. 2002. "The Study of Boundaries in the Social Sciences." *Annual Review of Sociology* 28(1):167–95. Retrieved August 10, 2016 (http://www.annualreviews.org/doi/10.1146/annurev.soc.28.110601.141107).

Lavizzo-Mourey, R. et al. 2001. "Attitudes and Beliefs About Exercise Among Elderly African Americans in an Urban Community." *Journal of the National Medical Association* 93(12):475–80.

Lee, I. M. and R. S. Paffenbarger. 2000. "Associations of Light, Moderate, and Vigorous Intensity Physical Activity with Longevity. the Harvard Alumni Health Study." *American Journal of Epidemiology* 151(3):293–99.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web." P. 342 in. ACM Press. Retrieved February 21, 2017 (http://portal.acm.org/citation.cfm?doid=1060745.1060797).

Mabry, Iris et al. 2003. "Physical Activity Attitudes of African American and White Adolescent Girls." *Ambulatory Pediatrics* 3(6):312–16.

Mathews, Anna E. et al. 2010. "Older Adults' Perceived Physical Activity Enablers and Barriers: A Multicultural Perspective." *Journal of Aging and Physical Activity* 18(2):119–40.

McCormick, T. H., H. Lee, N. Cesare, A. Shojaie, and E. S. Spiro. 2015. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing." *Sociological Methods & Research*. Retrieved August 28, 2016 (http://smr.sagepub.com/cgi/doi/10.1177/0049124115605339).

Mcguire, M., P. Hannan, D. Neumarksztainer, N. Cossrow, and M. Story. 2002. "Parental Correlates of Physical Activity in a Racially/Ethnically Diverse Adolescent Sample." *Journal of Adolescent Health* 30(4):253–61. Retrieved August 10, 2016 (http://linkinghub.

elsevier.com/retrieve/pii/S1054139X01003925).

Messias, Johnnatan, Lucas Schmidt, Ricardo Oliveira, and Fabrício Benevenuto. 2013. "You Followed My Bot! Transforming Robots into Influential Users in Twitter." *First Monday* 18(7). Retrieved February 15, 2017 (http://firstmonday.org/ojs/index.php/fm/article/view/4217).

Motalebi, Seyedeh Ameneh, Jamileh Amirzadeh Iranagh, Abbas Abdollahi, and Lim. 2014. "Applying of Theory of Planned Behavior to Promote Physical Activity and Exercise Behavior Among Older Adults." *Journal of Physical Education and Sport* 14(4):562–68.

Müller, Oliver, Iris Junglas, Jan vom Brocke, and Stefan Debortoli. 2016. "Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines." *European Journal of Information Systems* 25(4):289–302. Retrieved November 12, 2016 (http://link.springer.com/10.1057/ejis.2016.2).

Nguyen, Quynh C. et al. 2016. "Leveraging Geotagged Twitter Data to Examine Neighborhood Happiness, Diet, and Physical Activity." *Applied Geography* 73:77–88. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0143622816301394).

Nielsen, Finn \AArup. 2011. "A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs." *arXiv preprint arXiv:1103.2903*.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *ICWSM* 11(122-129):1–2.

Pak, Alexander and Patrick Paroubek. 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." in *Proceedings of the International Conference on Language Resources and Evaluation*. Valleta, Malta.

Pang, Bo and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2(1–2):1–135. Retrieved November 7, 2016 (http:

//www.nowpublishers.com/article/Details/INR-011).

Papacharissi, Zizi. 2012. "Without You, I'm Nothing: Performances of the Self on Twitter." *International Journal of Communication* 6. Retrieved (http://ijoc.org/index.php/ijoc/article/view/1484).

Park, Kunwoo, Ingmar Weber, Meeyoung Cha, and Chul Lee. 2016. "Persistent Sharing of Fitness App Status on Twitter." Pp. 183–93 in. ACM Press. Retrieved August 10, 2016 (http://dl.acm.org/citation.cfm?doid=2818048.2819921).

Pate, R. R. et al. 1995. "Physical Activity and Public Health. A Recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine." *JAMA* 273(5):402–7.

Paul, Michael and Mark Dredze. 2011. "You Are What You Tweet: Analyzing Twitter for Public Health"." Pp. 265–72 in *Proceedings of the Fifth International AAAI*. Retrieved (http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880).

Ping, Haoyue, Julia Stoyanovich, and Bill Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." P. 42 in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Popay, J., A. Rogers, and G. Williams. 1998. "Rationale and Standards for the Systematic Review of Qualitative Literature in Health Services Research." *Qualitative Health Research* 8(3):341–51.

Potts, Liza and Dave Jones. 2011. "Contextualizing Experiences: Tracing the Relationships Between People and Technologies in the Social Web." *Journal of Business and Technical Communication* 25(3):338–58.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved (https://www.R-project.

org/).

Ray, Rashawn. 2017. "Black People Don't Exercise in My Neighborhood: Perceived Racial Composition and Leisure-Time Physical Activity Among Middle Class Blacks and Whites." *Social Science Research* 66:42–57. Retrieved July 25, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0049089X17302764).

Rhue, Lauren and Jessica Clark. 2016. "Who Gets Started on Kickstarter? Racial Disparities in Crowdfunding Success." *SSRN Electronic Journal*. Retrieved February 17, 2017 (http://www.ssrn.com/abstract=2837042).

Rouse, M. 2013. "What Is Socialbot? - Definition from WhatIs.com." Retrieved February 15, 2017 (http://whatis.techtarget.com/definition/socialbot).

Scanfeld, Daniel, Vanessa Scanfeld, and Elaine L. Larson. 2010. "Dissemination of Health Information Through Social Networks: Twitter and Antibiotics." *American Journal of Infection Control* 38(3):182–88. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0196655310000349).

Schwarz, Susan Wile and Jason Peterson. 2010. *Adolescent Obesity in the United States: Facts for Policymakers*. New York: Columbia University. National Center for Children in Poverty, Mailman School of Public Health, Columbia University. Retrieved (http://hdl.handle.net/10022/AC:P:10726).

Shafahi, Mohammad, Leon Kempers, and Hamideh Afsarmanesh. 2016. "Phishing Through Social Bots on Twitter." Pp. 3703–12 in. IEEE. Retrieved February 15, 2017 (http://ieeexplore.ieee.org/document/7841038/).

Siddiqi, Z., J. A. Tiro, and K. Shuval. 2011. "Understanding Impediments and Enablers to Physical Activity Among African American Adults: A Systematic Review of Qualitative Studies." *Health Education Research* 26(6):1010–24. Retrieved August 10, 2016 (http:

//www.her.oxfordjournals.org/cgi/doi/10.1093/her/cyr068).

Skowron, Marcin, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. "Fusing Social Media Cues: Personality Prediction from Twitter and Instagram." Pp. 107–8 in. ACM Press. Retrieved May 30, 2017 (http://dl.acm.org/citation.cfm?doid=2872518.2889368).

Sloan, Luke et al. 2013. "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter." *Sociological Research Online* 18(3). Retrieved February 21, 2017 (http://www.socresonline.org.uk/18/3/7.html).

Smedley, Brian, Adrienne Stith, and Alan Nelson. 2003. *Unequal Treatment Confronting Racial and Ethnic Disparities in Healthcare.* Washington: National Academies Press. Retrieved August 10, 2016 (http://public.eblib.com/choice/publicfullrecord.aspx?p=3378816).

Smith, Aaron and Joanna Brenner. 2012. "Twitter Use 2012." *Pew Internet & American Life Project* 4.

Stephens, Thomas, David Jacobs Jr., and Craig White. 1985. "A Descriptive Epidemiology of Leisure-Time Physical Activity." *Public health reports* 100(2):147–58.

Streiner, David L., Geoffrey R. Norman, and John Cairney. 2015. *Health Measurement Scales: A Practical Guide to Their Development and Use.* Fifth edition. Oxford: Oxford University Press.

Tammelin, Tuija, Simo Näyhä, Andrew P. Hills, and Marjo Riitta Järvelin. 2003. "Adolescent Participation in Sports and Adult Physical Activity." *American Journal of Preventive Medicine* 24(1):22–28.

Tang, Jiliang and Huan Liu. 2012. "Unsupervised Feature Selection for Linked Social Media Data." P. 904 in. ACM Press. Retrieved May 30, 2017 (http://dl.acm.org/citation.

cfm?doid=2339530.2339673).

Trost, Stewart G., Neville Owen, Adrian E. Bauman, James F. Sallis, and Wendy Brown. 2002. "Correlates of Adults' Participation in Physical Activity: Review and Update." *Medicine and Science in Sports and Exercise* 34(12):1996–2001.

Tucker, Jared M., Gregory J. Welk, and Nicholas K. Beyler. 2011. "Physical Activity in U.S. Adults." *American Journal of Preventive Medicine* 40(4):454–61. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0749379711000122).

Turner-McGrievy, G. M. et al. 2013. "Comparison of Traditional Versus Mobile App Self-Monitoring of Physical Activity and Dietary Intake Among Overweight Adults Participating in an mHealth Weight Loss Program." *Journal of the American Medical Informatics Association* 20(3):513–18. Retrieved August 10, 2016 (http://jamia.oxfordjournals.org/cgi/doi/10.1136/amiajnl-2012-001510).

Unger, Jennifer B. et al. 2004. "Acculturation, Physical Activity, and Fast-Food Consumption Among Asian-American and Hispanic Adolescents." *Journal of Community Health* 29(6):467–81. Retrieved November 8, 2016 (http://link.springer.com/10.1007/s10900-004-3395-3).

Versey, H. Shellae. 2014. "Centering Perspectives on Black Women, Hair Politics, and Physical Activity." *American Journal of Public Health* 104(5):810–15. Retrieved August 10, 2016 (http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2013.301675).

Vicente, Marco, Fernando Batista, and Joao Paulo Carvalho. 2016. "Creating Extended Gender Labelled Datasets of Twitter Users." Pp. 690–702 in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, vol. 611, edited by J. P. Carvalho et al. Cham: Springer International Publishing. Retrieved February 17, 2017 (http://link.springer.com/10.1007/978-3-319-40581-0_56).

Voas, David. 2014. "Towards a Sociology of Attitudes." *Sociological Research Online* 19(1).

Retrieved October 14, 2016 (http://www.socresonline.org.uk/19/1/12.html).

Westerterp, Klaas R. and Guy Plasqui. 2004. "Physical Activity and Human Energy Expenditure." *Current Opinion in Clinical Nutrition and Metabolic Care* 7(6):607–13.

Wickham, Hadley. 2016. *Stringr: Simple, Consistent Wrappers for Common String Operations*. Retrieved (https://CRAN.R-project.org/package=stringr).

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2017. *Dplyr: A Grammar of Data Manipulation*. Retrieved (https://CRAN.R-project.org/package=dplyr).

Wilcox, S., C. Castro, A. C. King, R. Housemann, and R. C. Brownson. 2000. "Determinants of Leisure Time Physical Activity in Rural Compared with Urban Older and Ethnically Diverse Women in the United States." *Journal of Epidemiology and Community Health* 54(9):667–72.

Yadav, Daksha, Richa Singh, Mayank Vatsa, and Afzel Noore. 2014. "Recognizing Age-Separated Face Images: Humans and Machines" edited by E. Vul. *PLoS ONE* 9(12):e112234. Retrieved February 17, 2017 (http://dx.plos.org/10.1371/journal.pone.0112234).

Chapter 3

COMPARING TYPOLOGICAL APPROACHES TO FAMILY HOMELESSNESS

3.1 Introduction

Homelessness exposes individuals to multiple short- and long-term health-risks with life-long implications. Homeless episodes are associated with accelerated aging, mental health disorders, increased comorbid chronic diseases, and premature mortality (Fazel, Geddes, and Kushel 2014; Fischer and Breakey 1991; Morrison 2009; O'Connell 2005). Although homelessness is often considered in the context of single individuals (usually adult men), family homelessness has steadily increased since the 1980s and millions of children and adults (particularly women) are disadvantaged (Bassuk et al. 2014; Fertig and Reingold 2008; Metraux and Culhane 1999; Nunez and Fox 1999). Family homelessness is especially important because of intergenerational exposure to adverse life chances and the numerous public health resources (education, mental health, and housing resources[1]) involved with reducing family homelessness. Additionally, children involved with episodic homelessness are more vulnerable to developing behavioral problems, placement in foster care, under-performing in school, and adult homelessness (Bassuk and Rubin 1987; Culhane et al.

[1]Public health resources to reduce effects of homelessness: The No Child Behind Act provides funding for homeless school-aged children through the Department of Education (education) (Miller 2011), The Department of Health and Human Services (HHS) funds support services for individuals and families such as Temporary Assistance for Needy Families (TANF) and mental health services (mental health), and the Department of Housing and Urban development funds housing assistance programs for homeless families (Caton, Wilkins, and Anderson 2007; Culhane et al. 2007; Koh, Graham, and Glied 2011; Morgenstern et al. 2006).

2007; Grant et al. 2013; Masten et al. 2012; Rafferty and Shinn 1991; Shinn et al. 2008; Wood et al. 1990).

Effective homeless intervention program design and implementation rely on identifying groups of homeless service users to inform policy and pathways for exiting homelessness (Anderson 2001; Grigsby et al. 1990; Humphreys 1995; Mowbray, Bybee, and Cohen 1993). Prior research to identify and characterize homeless service users by demographic background relied on typologies created with cluster analysis on administrative or survey data to group individuals. Throughout the homeless typology literature, individuals and families are clustered in two methods. The first method, advocated by Kuhn and Culhane (1998)[2] relies on the theoretical belief that three types of homelessness service use patterns exist (Culhane et al. 2007; Fischer and Breakey 1991; Morse 1992; Shlay and Rossi 1992). The three types of homeless service users are the transitional homeless (typically a single, short episode of homelessness), episodic homeless (few episodes of short stays), and chronic homeless (multiple episodes of homelessness of lengthy duration). This typology uses "a two-dimensional mapping based upon number of episodes in a period versus number of cumulative shelter days in that same period" (Kuhn and Culhane 1998, pg. 215).

The second approach, advocated by Solarz and Bogat (1990), Morse, Calsyn, and Burger (1992), and Danseco and Holden (1998), does not start with a pre-determined group of clusters and attempts to group homeless service users by household demographic characteristics. Demographic characteristics in these studies include number of moves, substance abuse history, mental health status, employment history, amount of children, etc. (See Appendix B for demographic variables used in previous studies). Recently, both approaches to homelessness typologies have been extended to include families as well (Culhane et al. 2007; Culhane, Park, and Metraux 2011; Danseco and Holden 1998;

[2]Arce and Vergare (1984) did develop a three cluster homeless typology for the mentally ill before Kuhn and Culhane (1998) applied a theoretically defined clustering approach to homeless individuals.

Kuhn and Culhane 1998). Homelessness research has used survey and administrative data to analyze homelessness typologies. Survey data is challenging because respondents can refuse interviews or conceal homelessness experiences, short-term episodes of homelessness may be missed, and surveys may miss geographic differences (e.g., urban vs. suburban) in homelessness prevalence (Link et al. 1994). Alternatively, administrative homeless data may depict a more accurate homeless census, especially for homeless families because of the growth of Housing and Urban Development family-focused housing opportunities such as transitional housing (Burt and others 2006; O'Connell 2003). Administrative data is derived from client demographic information at the point of entry into a service provider's system, and previous research used these data to identify types of individual and family homeless service users with traditional cluster analysis (Culhane et al. 2007; Kuhn and Culhane 1998). However, administrative data has veracity concerns that limit analyses. New clustering algorithms with machine learning cryptography methods can further leverage administrative data to identify groups of homeless services users and inform both the census of homeless families and depict the diverse needs homeless families face (all while maintaining data privacy and consistency).

This paper's purpose is to directly compare clustering approaches and create a discourse between theoretically-driven typologies and bottom-up approaches to understanding homeless populations. With the aid of three types of synthetic administrative data from a county comprising one major metropolitan area, I examine homeless family typologies by using finite normal mixture models to explore homelessness family typologies. Analysis reveal that in each synthesized data type, the approach used determines the amount of clusters derived. For instance, clustering on homeless experience broadly captures the different groups of users and needs consistent with theoretically-driven typology literature. Additionally, results show that even the highest mode of data privacy (random synthesis) is able to produce comparable results to less randomized sampling strategies and supports growing research meaningful conclusions can be derived from anonymized (or

inaccurate) data produced with cryptographic sampling methods (Reiter 2002). Lastly, this study contributes to family homelessness literature by directly comparing typologies to group homeless families and discussing the limitations and challenges introduced from using synthetic administrative data and emerging methodology to understand homelessness resource utilization.

3.2 Background

Homelessness in the US

The existence of multiple definitions of homelessness complicates homelessness research. Common perceptions of homelessness focus on individuals literally without housing, but this definition is quickly strained. Lee, Tyler, and Wright (2010) describes disagreement in homelessness scholarship about counting squatters, persons leaving in cheap hotels, individuals in domestic violence facilities, residential treatment programs, and mentally-ill focused transitional housing. Although there is disagreement about the individuals that constitute the homeless population, researchers have more agreement about the causes and consequences of homelessness. Traditionally, homelessness is considered the byproduct of macro-level (e.g., housing affordability, economic and demographic changes, and policy changes[3]) and micro-level factors (e.g., criminal justice interactions, mental health issues, substance abuse[4]). Literature on homelessness has revealed that the cyclicity of homelessness is better understood in terms of homelessness frequency and duration instead of the dichotomous homeless/not homeless. Thus, theoretically driven homelessness scholars have focused on three types of homelessness and understanding the

[3](Sudden and long-term) poverty has a central role to most homelessness experiences. Declines in social support services such as TANF funding are important causes of family homelessness (Page and Nooe 2002). See: Hertzberg (1992) for a review on poverty and homelessness.

[4]Substance use and mental health issues are most common health issues for homeless adults(Aiemagno et al. 1996; Bassuk 1984). See Draine et al. (2002) and Galea and Vlahov (2002) for discussion on relationship between criminal justice system, mental illness, and homelessness.

demographic similarity between homeless service users with episodic, chronic, or transitional typologies (Culhane et al. 2007). These typologies are meant to improve homeless interventions by associating shelter utilization with homeless population characteristics.

3.2.1 Typologies of Homelessness

Historically, homelessness typologies started with single adults, but have grown to focus on subgroups such as runaway youth and families as well. Homeless family typologies focus on homeless service utilization, household demographic information, and (child) behavioral data either in isolation or together. Key differences between typologies reflect the clustering method (homeless experience or demographic background) as well as using prospective or retrospective data and survey or administrative data.

Individual typologies

Research using individual background characteristics to understand homelessness used clustering algorithms that considered a wide array of demographic characteristics and personal history. Solarz and Bogat (1990) focused on the effects of criminality, psychiatric background, history of transient behavior, and previous criminal victimization to derive seven groups of homeless individuals (pg. 86-87). Similarly, in explicitly rejecting top-down theoretical descriptions of homelessness, Morse et al. (1992) concludes that four clusters capture the subgroups of homeless users when clustering homeless individuals by psychopathology, alcoholism, social support, socioeconomic status, and health status. Researchers that have investigated mental health status, social networks, employment histories and other individual background characters tend to identify four-cluster models (e.g., recently dislocated, vulnerable, outsiders, and the prolonged homeless. See Humphreys (1995), Grigsby et al. (1990), and Mowbray et al. (1993) for four-cluster typologies)[5].

[5](1) recently dislocated: small social networks and mild mental health problems; (2) vulnerable: long duration of homelessness, sparse social networks, and more severe mental health problems; (3) outsiders:

Later clustering studies attempted to improve on earlier atheoretical models by removing demographic characteristics from clustering models, and favoring clustering individuals along their homeless experiences to separate the causes and consequences of homelessness (Kuhn and Culhane 1998). Kuhn and Culhane (1998) claims that homelessness literature suggests building "a theoretically grounded model based only on homeless experience, attempting to confirm that the demographic, socioeconomic, and treatment backgrounds of the clusters are distinct from each other and reflect the expected characteristics associated with membership in a given cluster" (209-210). Once again, the three theoretically driven clusters that emerge from the literature reviewed by Kuhn and Culhane (1998) and these authors' own research are transitionally homeless, episodically homeless, and chronically homeless (210-211). Transitionally homeless individuals are defined by typically, one short shelter experience (one short period stay). Episodically homeless individuals are often younger individuals that experience multiple shelter stays with shorter duration (many episodes of fewer than a month). Lastly, the chronically homeless tend to be older populations that experience years-long homelessness and service use ("skid-row homeless", pg. 211). With homeless administrative data from New York (1986-1995) and Philadelphia (1991-1995), these authors find that 81% of observations were transitionally homeless, 9% were episodically homeless, and 10% were chronically homeless (pg. 219 - Table 4).

Family typologies

Family homelessness typologies emerged from a need to characterize the mechanisms by which homelessness affects children and verify the homeless experience typology developed from the experiences of single adults. Researchers on homeless individuals have suggested that homeless service typologies may not correspond exactly to homeless fami-

---

very similar to the vulnerable, but with more robust social networks and less severe mental health problems (4) prolonged: extensive homeless experience (more than five years), sparse networks, and moderately dysfunctional (Grigsby et al. 1990).

lies although individual typologies have been used as a reference point (Bassuk, T Volk, and Olivet 2010; Benjaminsen and Andrade 2015; Culhane et al. 2007, 2011; Danseco and Holden 1998; Kuhn and Culhane 1998). Early family homelessness typologies evaluated the relationship between shelter service use and child outcomes. Danseco and Holden (1998) focused on developing a typology to determine the parenting characteristics (e.g., stress, employment, substance abuse, etc.) that differentiated child outcomes in homeless families. These authors constructed three clusters of families (at-risk, getting-by, resilient)[6] that predicted child education and behavioral outcomes suggesting the need to understand behavioral and historical variability within families when designing intervention programs. Burt (2001) used multiple two-dimensional modelling approaches with data from the 1996 National Survey of Homeless Assistance Providers and Clients (NSHAPC) and proprietary Urban Institute. These authors found three clusters of homeless service use amongst homeless families (episodic, chronic, and crisis oriented). However, the size and fit of these clusters did not parallel previous two-dimensional clustering research from Kuhn and Culhane (1998).

Culhane et al. (2007) attempt to verify the individual homelessness resource utilization typology for families and concluded that individual homelessness typology broadly applies to families with some notable differences. This analysis used multi-year administrative data from Philadelphia, New York City, Columbus (OH) and the state of Massachusetts and found three types of usage clusters (transitional, episodic, and chronic) existed with families, but the relative proportion of each cluster was different from single adults and strong demographic differences between the three clusters were less apparent

[6](1) at-risk: high levels of parenting stress and parenting characteristics associated with negative child outcomes; (2) resilient: lowest parenting stress and parenting characteristics that were protective for child outcomes; (3) children from these families had several outcomes associated with lower-levels of socio-economic status such as academic underachievement. The from Baltimore City Families In Transition program that was administered from October 1992 to March 1992 (n = 180 families, 348 children; see page 160 for more descriptive statistics of this data) (Danseco and Holden 1998).

than in single adults. This study makes overt ties with the language and characterization of individual homeless typologies through the definitions of transitional, episodic, and chronically homeless. Similar to Kuhn and Culhane (1998), Culhane et al. (2007) define transitionally homeless families as families that experience typically one short shelter stay, episodic homeless families experience multiple shelter stays with shorter duration, and chronically homeless average less than two shelter stays of long periods. Another study by Culhane, Kuhn, and colleagues studied families current use of inpatient behavioral health and child welfare placement services with linked, personally-identifiable data from multiple administrative databases in Philadelphia (Culhane et al. 2011). These authors used the three homelessness service use experience clusters (transitional, episodic, and chronic) reported in their previous work. Analyses found that 71% of families were transitionally homeless, 8% were episodically homeless and 21% were chronically homeless (819).

For the purpose of this study, comparing individual homeless experience typologies with nascent family homelessness is most important. Gleason, Barile, and Baker (2017) analyzed statewide homelessness service provider administrative data on individuals and families (Hawaii) with latent class growth analysis (a structural equations modeling method) and determined a four-cluster resource utilization typology (low service, typical transitional, atypical transitional, and potential chronic service use) that departs from the three cluster typology consistently used in the literature[7]. Homeless family typologies have clarified how child welfare is influenced by various established homeless experiences,

[7](1) low service: low usage across three homelessness services (emergency, transitional, and outreach services) , (2) typical transitional: initially high-levels of transitional shelter use and eventual exit from homeless cycle , (3) atypical transitional,: initially low-levels of transitional shelter, but higher use of emergency shelter use than the typical transitional (4) potential chronic service use: high-levels of emergency shelter and outreach services (Gleason et al. 2017). Benjaminsen and Andrade (2015) represent a more recent study that followed Kuhn and Culhane (1998) in applying a three cluster service use approach in a comparison between Denmark and the United States.

but have not investigated the full range of family homeless experiences. Exploring the stability of the (individual) three-cluster usage models is needed to understand how family needs relate to existing program goals.

Determining family homelessness typologies with administrative data

Traditionally, homelessness service use is measured with survey and administrative data. These data sources have their own strengths and limitations, but the relative strengths of administrative data outweigh its limitations. Survey data use sampling methods that allow individuals from the population and equal chance of representation and are flexible enough to measure specific research questions. However, survey methods rely on retrospective recall and have with respondent refusal and survey reliability that are key concerns with homeless populations. Administrative data is limited by data accuracy, missingness, and the general difficulty translating administrative data into research concepts. Administrative data document individuals' employment status, substance abuse as well as information about children such as child's insurance status and mental and physical well-being Additionally, administrative data has privacy concerns that are not easily controlled; preserving client privacy while maintaining a detailed, informative representation of the data is another hurdle for studies with administrative data. Despite the limitations, administrative data reflect service usage patterns which is ideal for understanding homelessness resource utilization.

Administrative data on homelessness is generally created at the point an individual begins a relationship with a homeless service provider (e.g., transitional housing, rental assistance, rehabilitation services, etc.). Homeless service providers that receive funding from the U.S. Department of Housing and Urban Development (HUD) as well as the Department of Health and Human Services (HHS) must comply with data mandates to use an electronic record system known as the Homeless Management Information System (HMIS). HMIS standardizes questions providers ask clients such as substance use, receipt of state or federal-aid (e.g., TANF) along with demographic characteristics.

3.3 Data and Methods

The limitations of administrative data can be directly addressed by methodological advances in producing synthetic data and clustering. Rubin (1993) proposed creating synthetic data sets to address confidentiality concerns for secure data. Synthetic data sets are designed to ignore sampling schema, use multiple imputation, and most importantly for this study incorporate intentionally inaccurate (e.g., not the actual collected value in the original data set) data for some variables. Reiter (2002) and others have found that valid inferences can be gained from synthetic data. Given the concerns with data and respondent reliability in HMIS, using intentionally accurate data to model relationships builds robustness. Models that project forward with HMIS data should be sensitive to unknown, inaccurate data instead of face-value acceptance of the HMIS data quality.

Improvements in clustering methods may benefit the identification of resource utilization groups. Theoretically driven homelessness research has confirmed three types of homeless experiences through clustering methods such as k-means clustering and nearest centroid (cluster) that assign cases to clusters by minimizing the (Euclidean) distance between a cluster's mean and each case (Anderberg 1973; MacQueen and others 1967). However, advances in model-based clustering have allowed researchers to use a broader array of variable types to predict groups (Anderlucci and Hennig 2014; Hennig and Liao 2010). Mixture models are a model-based clustering approach tested with social phenomena such as social stratification (Hennig and Liao 2010).

This study uses a multivariate finite normal mixture models with two approaches in homelessness research that leverage family homeless experience factors including program duration and program count. Program duration measures the total amount of days that a family spent accessing homeless services across the programs that family was enrolled in and programs count reflect the amount of homeless episodes experienced by that same family. The outcome variables were standardized by using z-scores to simplify the models and reduce the parameter space. Z-scores represent the number of standard deviations

(dispersion) an individual observation is above or below the mean group observation. Furthermore this study, leverages multiple synthetic data sets, representative of one greater metropolitan area with finite normal mixture models to determine typologies of homeless family service use. HMIS were collected from multiple service providers in HUD and HHS funded programs such as transitional housing, emergency shelters, rapid-rehousing, support services only, etc. Each observation represents a unique family's homeless experience (duration, programs used, and (if an exit interview was administered) respondent's living situation at the time of program exit). The synthesized independent county and synthesized random county are similar in size (these synthesized counties are 120% the size of synthesized correlated county) and these jurisdictions are based on longitudinal program data from 1993 to 2015.

Synthetic data

Synthetic data is used to maintain the privacy of individuals in the HMIS data and introduce data uncertainty that mimics real word data entry and respondent accuracy limitations. A major challenge with synthetic data is that realistic enough data potentially violate individuals' privacy (for example, by saying things about their neighborhood). On the other hand, if the data is obfuscated, it might become less useful for drawing any conclusions. Advances in cryptographic methods such as differential privacy have demonstrated the ability to preserve client privacy, while maintaining a connection to the underlying data (Chawla et al. 2005; Dwork 2006, 2007). Given difficulties with retroactive use of data (data entry, data quality, etc.) cryptographic and machine learning methods that are novel to homelessness research could clarify inferences (Metraux and Tseng 2017). Furthermore, algorithmic bias with sensitive data requires HMIS funded organizations adopt a unique workflow because these HMIS programs can't give analysts private data to develop tools. Instead, homelessness service providers can leverage differential privacy by masking the "true" value of any respondent covariate with noise and missingness and still allow analysts to develop tools that can then be applied to real data (Barrientos et al.

2017). Respect for data sharing agreements (data sensitivity) along real-world limitations with data entry (data accuracy) make HMIS data ideal for synthetic analysis.

The synthetic data sets were created by using the open source software DataSynthesizer[8]. DataSynthesizer can synthesize data in three different modes that add varying degrees of differential privacy (noise) (Ping, Stoyanovich, and Howe 2017). The basic approach for each mode is the same (create a distribution for each variable and samples that distribution); however, the degree of differential privacy and missing rates used to further anonymize the data change[9]. The software for creating synthetic data has a noise parameter that is equivalent to the minimum change in the correlations by removing (perturbing) any tuple of data at random. Missingness is determined by the synthetic mode utilized. In the random mode, ideal for the most sensitive data, the DataSynthesizer generates type consistent random variables, adding the highest degree of differential privacy. The correlated attribute mode uses a Bayesian network to calculate the relationship between variables and in cases where calculating a Bayesian network is too computationally expensive Ping et al. (2017) suggest using the independent attribute to sample from a noise-added distribution of the underlying data. After synthesizing a single county's HMIS (and sampling and untold percentage of this county[10]), the synthesized data was synthesized again to enhance data privacy and align with data sharing agreements. Thus, the final data set is a twice synthesized representation of HMIS data and is a definitive abstraction of any county HMIS data.

Clustering

At the point of entry into a homeless service provider, individuals are given a unique

[8] `https://github.com/DataResponsibly/DataSynthesizer`

[9] The synthetic data uses a schema model instead of a generative model of social processes that create the data. The key distinction between schema model and generative model is the schema model is focused on building tools to explain the data while the generative model focuses on theories to explain the data.

[10] Due to the sensitivity of the data, the amount sample will not be revealed

identifier and families are usually grouped with a family identifier. This approach has methodological limitations based on accuracy and continuity. First, data entry errors mean that individuals within a family are not always assigned the same family identifier, implying some family units (within individual service providers) are misidentified families. Secondly, families that utilize more than one service will most likely be given different family identifiers with every provider so the continuity of families is not ensured[11]. Furthermore, while the data is longitudinal, it does not represent the historical sequence of service use important for understanding trajectories. Families were created by using a clustering approach developed by Rokem and Hazelton (2016) that relies on the family identifier and time individuals registered for a homeless service[12]. Finite normal mixture models were developed to analyze change over time and understand how covariates inform longitudinal processes (Figueiredo and Jain 2002; Fraley and Raftery 2006; McLachlan and Peel 2004; Proust-Lima, Philipps, and Liquet 2015; Rasmussen 2000).

Analytical strategy

To bridge homelessness typology research, this study employs two modeling approaches. The first approach is based on pre-defined three cluster typology advanced by Kuhn and Culhane (1998) and others and uses a multivariate finite normal mixture models with two dependent variables (standardized duration of homeless episodes and amount of homelessness episodes[13]) that correspond to the resource utilization typology consistent in homelessness research. The second approach follows Danseco and Holden (1998) and attempts to determine the ideal amount of homeless clusters with the demo-

[11]The HMIS data used for this study did not have identifiable individual data (e.g., Social Security numbers) to consistently group individuals into families. HMIS data with unique personal identifiers may not have as many issues with accuracy and continuity for identifying families.

[12]`https://github.com/uwescience-bmgf-hmis/puget`

[13]The outcome variables were standardized to simplify the models. Program duration had a range of 7141 while program counts had a range of 7. All code and data will be available at `https://github.com/kpolimis`

graphic variables about families. The data correspond to longitudinal observations with at least an entry and exit interview. The combination of these two interviews are the basis of the two-dimensional approach's focal variables (duration of services, program history). The family background variables are collected at the time a client enters into a service provider's system (entry interview). The continuous variables in the family background approach approach are: amount of adults in the household and number of children and the binary variables: household structure (one or two parent), employment status of adults in the household, current substance abuse, veteran status, parents' physical disabilities, parents' mental disabilities, and the receipt of any federal or state benefits (e.g., Medicare/Medicaid, TANF, State Children's Health Insurance Program (SCHIP)).

In addition to comparing the two-dimensional and family background approach to generating homeless family typologies, this study also uses an atheoretical model that finds the best available clusters from a range of 1 to 20 clusters (this is the range available in the 'mclust' package). The findings and discussion focus on the comparing the clusters reviewed in the literature, either three or four clusters of homeless families and clusters; the atheoretical clusters are also analyzed separately (Danseco and Holden 1998; Gleason et al. 2017; Kuhn and Culhane 1998). The analysis was completed R software (R Core Team 2016) and relied heavily on the `mclust` package for cluster analysis (Fraley and Raftery 2006; Scrucca et al. 2016).

Exclusionary Criteria

Observations were excluded for missingness on any demographic variable or homelessness outcome. In the synthesized random county, 0% of the data were dropped due to missingness. Synthesized independent county and synthesized correlated county lost 0% and 19%, respectively.

3.4 Findings

The three class resource utilization typology (transitional, episodic, chronic) that dominate the literature was tested by specifying 3 and 4 class (cluster) models as well as an atheoretical model that tested 1 to 20 clusters. The models are compared with the Bayesian Information Criteria (BIC), a function that incorporates each model's maximized log-likelihood, the data dimensions, and number of parameters used in cluster estimation (Dean and Raftery 2010; Proust-Lima et al. 2015). "The BIC is the value of the maximized log-likelihood with a penalty on the number of model parameters, and allows comparison of models with differing parameterizations and/or differing numbers of clusters. In general the larger the value of the BIC, the stronger the evidence for the model and number of clusters" (Fraley, Raftery, and Scrucca 2012, pg. 19). Model BIC are available in Appendix B (Tables B.1 and B.2). Multiple comparisons are made in this discussion to unite discussions about homeless family typologies. First, the two dimensional and family background models of clustering are compared across all synthetic modes of data. Secondly, three and four cluster models are compared within these clustering strategies because these are the most common clusters found in previous literature (literature comparison). Lastly, the best available clusters from the atheoretical model is presented (atheoretical comparison). Findings for the two-dimensional approach are presented first followed by findings for the family background approach.

The summary statistics for each synthesized county data are available in Table B.3 (Appendix B). In the summary statistics, the components of the DataSynthesizer can clearly be seen. For instance, the random synthesized data has by far the highest mean duration for homelessness and programs used (mean duration: 3568.15 is 10 times the other synthesized data modes, 1256.63 (independent) and 451.93 (correlated)). As these measures are the key clustering variables for the two-dimensional approach, the clustering comparison across modes is also likely to reflect the sampling differences. Additionally, key demographic differences are observed between the three modes of synthesized data.

The synthesized random county has near equal proportions of the three types of households (single men, single women, and couples are proportionally 0.25, 0.25, 0.25 of the data). The equal representation of these groups is related to the random variable sampling. Conversely, 0.77 of the independent synthesized county and 0.74 of the correlated synthesized county are represented by single females.

3.4.1 Literature Driven Approaches

3.4.2 Two-dimensional Clustering

In the literature comparison of the two dimensional approach, four groups were found as the best characterization of the data in the random, independent, and correlated attribute mode for synthetic data (Appendix B). Four clusters of homeless service use is inconsistent with Culhane et al. (2007) an early study applying the two-dimensional approach to determine homeless family typology. Culhane et al. (2007), the first to verify the individual typology with family homelessness (with a two-dimensional clustering approach), found that transitional homeless represented between 72% and 80% of the homeless population, the episodic homeless accounted for 18%-22%, and the chronically homeless represented between 2% and 8% in the four jurisdictions they analyzed.

Gleason et al. (2017) also attempt to combine the two-dimensional approach and family background approach by starting with a two-dimensional model and adding (groups of) demographic characteristics. Their model includes type of homelessness service used (e.g., emergency shelter, transitional shelter, and outreach services) and the inclusion of homelessness service is an important dimension used to determine their model's classes (review footnote six for a description of the service use classes their model determined) and their conclusions are not necessarily comparable to this analysis. Besides differences in the synthetic data used here and the regional data from multiple metropolitan areas, modeling strategies such as variable selection may also drive differences in findings.

Synthesized data - Random Attribute Mode

There does not appear to be significant demographic differences across these clusters due to random variable sampling, although the clusters differ in size (Table B.4). While the transitionally homeless are the largest group in individual and family typology studies, since there are four groups instead of the traditional three, it appears that the transitionally homeless have "split" their defining characteristics. One group that can be considered transitionally homeless is Cluster 3 because it contains the lowest average program duration days (1658.05 days); however, this cluster does not have the second defining feature of the transitionally homeless, fewest programs used (atypical programs). Transitionally homeless with atypical programs were 32.2% of the data. Cluster 1 has the fewest average programs although the mean duration days of those individuals resembles the group that is closest to traditional literature's episodically homeless group, Cluster 2 and 4. Cluster 1, transitionally homeless with atypical days contained 17.88% observations. The two clusters closest to the chronically homeless, Cluster 2 and 4 combined multiple characteristics of the chronically homeless as described in the literature (Culhane et al. 2007). Cluster 4, had the chronic characteristic of highest average amount of programs used (6.76 programs) while Cluster 2 had stays that averaged the longest average duration (5567.76 days). Cluster 2 and 4 account for 30.7% and 19.23% of data; respectively. The appearance of multiple transitional and chronic clusters as well as other differences with the traditionally clustered family groups separate the findings from randomly synthesized data from previous literature (recall the random sampling strategy and relatively even distribution). Once again, the use of random variable sampling make the results from the random synthesized data particularly difficult to interpret.

Synthesized data - Independent Attribute Mode

The model BIC for the independent synthesized data suggest that a four cluster model is best. The largest group was an atypically transitional group (Table B.5). Two-dimensional models suggest that the transitionally homeless should be the largest group. The independent synthetic data appear to have a traditionally transitional cluster as

well as two hybrid transitional clusters. Cluster 2 (the largest group) can be considered traditionally transitional because it contains the lowest average programs used (1 program) and program duration (188.87 days). Traditionally transitional homeless families represent 41.13% of the data. Two clusters, clusters 3 and 4 can be considered atypical transitional because of their low average program use. Cluster 3 (7.74% of the data) is near identical to the traditional transitional cluster in average programs used (2) and program duration (209.03 days). The largest difference between clusters 2 and 3 is slightly greater substance use for the atypical transitional third cluster (0.25 compared to 0.36). The second atypical transitional cluster is Cluster 4 (17.69% of observations) because this cluster also averages the fewest programs (1), but the average program duration (719.66 days) is considerably higher than two other transitional clusters. Cluster 1 appears to confirm to the chronically homeless, a group of families with long-term stays (average stay averaged 3108.42 days), although these stays were not numerous (2.35 programs used on average). Chronically homeless accounted for 33.43% of the observations. There are no significant demographic differences across any of the clusters besides substance use discussed above.

Synthesized data - Correlated Attribute Mode

The model BIC for the correlated synthesized data also indicated that a four cluster model is best. Once again, the largest group was a traditionally transitional group. In this mode of synthesized data (Table B.6), both standards in previous literature for transitional status was met. Cluster 1 averaged lowest average program used (1.01) and fewest average program days (144.76). The transitionally homeless represented 43.49% of the data. Similar to the synthetic independent data, two other transitional-like homeless clusters were observed. The first atypical transitional cluster, cluster 3, used resources on par with the traditionally transitional group (1.02 programs used on average), but the average program stay of this cluster was the third greatest overall (averaged 589.47 days). The second atypical transitional cluster, cluster 4, displayed the second lowest average days in a program (averaged 235 days per stay), but the third greatest average

programs used (2.27 programs used). Cluster 3 represented 33.25% of the data while cluster 4 accounted for 8.96% of observations. Lastly, the second cluster can be described as chronically homeless because these families averaged the longest average stay (1205.73 days) and frequent use of services (3.39 programs used on average); this cluster was the appropriate classification for 14.3% of the homeless families. There is more racial diversity in the clusters than observed in either the independent or random synthetic data sets.

3.4.3 Family Background Clustering

Clustering homelessness families by demographic background revealed greater inter-cluster demographic differences than clustering on homelessness experiences. Similar to the two-dimensional approach, three and four-cluster models are compared with the atheoretical modelling section addressing the objectively best fitting grouping of data. Additionally, in all three modes of synthesized data four-cluster models were considered the best by BIC (Table B.2). Background clustering approaches tend to derive four clusters on individual and family data (Gleason et al. 2017; Grigsby et al. 1990; Humphreys 1995; Morse et al. 1992; Mowbray et al. 1993). While the clusters are demographically distinct, the homeless experiences do not vary significantly (this is the reverse of the two-dimensional approach where homeless experiences did vary significantly, but the clusters were demographically similar).

Synthesized data - Random Attribute Mode

The model BIC for the random attribute mode of synthesized data show that a four cluster model is best. The homeless experiences across the clusters are related to variations in household racial background, substance abuse, and the average amount of children. The summary statistics for each cluster are available in Table B.7.

There are generally small household structure differences (exception: amount of children) across the clusters although racial distinctions are prominent. Two clusters are near equal-sized and larger than the remaining two clusters. Cluster 1 and 3 contain

1502 and 1506 observations, respectively. Contrastingly, cluster two has 758 observations while cluster 4 has 1502 families. Cluster 1 contains 32.2% of the data and is composed of 26% single women, 25% single men, and 23% couples. However, 65% of the cluster is composed of racial minority households. Racial minorities are well represented in all clusters, especially cluster 3 where they make up 100% of observations with non-missing race. However, with the exception of cluster 4, the representation is not equal from each minority group. For instance, Blacks are not present in the first cluster and Pacific-Islanders are not represented in the third cluster. The second cluster is 16.25% of all observations and 84 of these families are racial minorities. The household structure for this cluster is very similar to cluster 1 where there is slightly less racial diversity: 19% single women, 21% single men, and 30% couples. Besides racial differences across clusters, other demographic variation include the amount of children. The amount of children in the household is the largest demographic difference across these four clusters. Cluster 4 had the fewest average children (9.52) while cluster 2 averaged the greatest (1.39).

Synthesized data - Independent Attribute Mode

The model BIC for the independent attribute mode of synthesized data indicate that a four-cluster model is preferred to three clusters. Once again, the homeless experiences across clusters are not very different. The mean program duration and programs used for each cluster are below in Table 3.1 (extended summary statistics for each cluster in Table B.8).

Table 3.1: Independent Synthesis - Homeless Experience

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	1222	1283	1240	1270
mean programs (count)	1.6	1.55	1.51	1.51

Three of the four clusters are similar in size with one cluster much smaller than the rest. There are significant racial and household structure differences across clusters. Cluster 1 represents the smallest cluster and is 9.33% of the data. Within this cluster, 18% of the homeless families are couples, 78% are single females, and 5% are single males. 67% of the families in this cluster are racial minorities. Cluster 2 is dominated by black coupled households and is composed of 50% single women, 10% single men, and 40% couples. 100% of the families in cluster 2 are racial minorities and this cluster comprises 26.93% of all observations. The last racial minority dominated cluster is Cluster 4 where 98% of observations are single women, 1% are single men, and 1% are couples. The families in this cluster represent 34.84% of the data and 40% of the families in Cluster 3 are racial minorities. In sharp contrast to Clusters 1, 2, and 4, Cluster 3 is 100% White. Additionally, 77% of Cluster 3 is single women, 4% single men, and 18% couples.

The only other significant demographic difference occurred with substance abuse, a behavior most prevalent in Cluster 1 and almost unreported in other clusters. The first cluster is primarily composed of single females from racial minority backgrounds.

Synthesized data - Correlated Attribute Mode

The model BIC for the correlated synthesized data indicate that a four cluster model is best. The homeless experiences across clusters are more varied than other data types although many clusters have a similar experience. The mean program duration and programs used for each cluster are below in Table 3.2 (extended summary statistics in Table B.9).

Table 3.2: Correlated Synthesis - Homeless Experience

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	418.5	507	479.2	438.3
mean programs (count)	1.43	1.52	1.5	1.45

The data displays dramatic racial and household structure differences across clusters. One cluster is significantly larger than the rest while two clusters are similar in size and the last cluster is 12% the size of the next smallest cluster. The first mid-sized cluster is Cluster 1 and this group represents 23.48% of the data and is 75% single female with 7% of these women from racial minority backgrounds. Cluster 1 differs markedly in racial and demographic composition with the remaining groups. Cluster 2 is the second mid-sized cluster and is composed of 75% single women, 8% single men, and 17% couples. Cluster 2 accounts for 25.75% of all observations and 88% of these families are White (the remaining 12% of families are Pacific-Islander). Cluster 3 is the smallest cluster and most diverse racially. 59% of observations are single women, 22% are single men, and 19% are couples. The families in this cluster represent 2.71% of the data and 20 of the families in Cluster 3 are White. The largest cluster, cluster 4, accounts for 48.06% of all observations and 100% of these families are racial minorities.

3.5 Atheoretical Clustering

After comparing three- and four-cluster models with the two-dimensional and family background approach, atheoretical models were used to determine the best grouping for the data. Atheoretical models were tested within the two-dimensional and family background approach and revealed considerably more clusters (8-12) than discussed in either literature (for exceptions see: Solarz and Bogat (1990) that found seven clusters). Atheoretical models also combined the two-dimensional and family background approach to understand the joint influence of homeless experience and family background. The results from the atheoretical models that combine both clustering approaches and do not restrict the amount of clusters a priori are even more striking than atheoretical models tested within the two-dimensional or family background approach. The clusters range from a low of 12 with the random synthetic data to a high of 20 clusters in the correlated data (the clustering software has an upper-bound of 20 clusters, all atheoretical data and tables will be available at `https://github.com/kpolimis`). While the random synthetic data does not

reveal much variation between homeless experience (mean programs used and program duration), compelling demographic differences such as household structure, substance use, and the amount of children emerge. The independent and correlated data also illustrate important, but different demographic variation as well as significant differences in homeless experience. For instance, the range of the program duration spanned approximately 3400 days and 6 programs in the correlated synthetic data (the range was approximately 4800 days and 5 programs in the independent synthetic data). Moreover, the demographic differences observed in these synthetic data were different from the random synthetic data. In the independent attribute mode, parental employment and disability were important features while the correlated mode indicated veteran status varied across clusters. A key takeaway from the atheoretical models that combine the family background and two-dimensional clustering approach is the presence of numerous clusters and demographic background stratifying. Closer examination of these clusters suggest that they can be condensed to capture broad distinctions such as household, racial, and homeless experience differences without creating a small group for every demographic combination. For instance, instead of having two identical clusters that differ along the lines of one is mostly Black and the other non-Black minorities, these clusters could be condensed into a single group.

3.6 Limitations

This study has several limitations related to the longitudinal nature of the data, available demographic features, and data quality. Clients homeless experiences were not sequentially recorded and some of the homeless family experience could be obscured limiting the ability to investigate the role of homelessness trajectories. Furthermore, this analysis did not have or consider child and criminal justice interactions that could have further differentiated homeless demographic groups. Lastly, data fidelity and geographic limitation temper generalizability and suggest more research should be done to assert the primacy of the three cluster homelessness usage typology. Future research could compare synthe-

sized county data in several ways. First, analysis of greater time periods of data with more demographic features could further understand family differences in homelessness service usage. Secondly, within multiple synthetic datasets, various clustering approaches could be used for cross validation. For instance, group models could be compared to centroid and neural models to determine how much model selection impacts the identification of clusters. Lastly, this project can connect homeless typologies to homeless service usage to understand how demographic variations influence successful exits from homelessness.

3.7 Conclusion

The type of synthetic data used in this study doesn't allow for a better understanding of homelessness per se because explaining the factors that produce homeless families is secondary to analyzing the differences between these same families. To that end, this study compared three modelling approaches with and found larger clusters of homeless families than typically described in theoretically driven literature. When choosing between a three and four-cluster model, the four-cluster model dominated both the two-dimensional and family background approaches in the random, independent, and correlated attribute mode.

Family homeless typologies produced with multivariate finite normal mixture models are largely incompatible with the three types of homelessness service use (transitional, episodic, and chronic) in all three synthesized modes of HMIS county data analyzed. Initially, the randomly synthesized county appears very different from the other modes of synthetic data (the average program duration is 10 times the duration of the other synthetic data modes with significantly more homeless programs), but the substantive conclusions and clusters derived are similar.

Contributions from this study suggest that homelessness needs are more varied than the three cluster homelessness service use typology (episodic, transitional, and chronic) that is persistent in studies of individual homelessness and emerging in studies of family

homelessness. Although the data were synthesized and not representative of any one county, the finding that homelessness service has more clusters than typically discussed demonstrates the variation in the homeless family population and the potential need to redefine service delivery. The introduction of machine learning cryptography for data sensitivity (and potential inaccuracies) in combination with the latent mixture class model could drive these results herein that diverge from the existing literature and future research could clarify the degree novel methods influence findings versus the existence of distinct family homeless populations. A key finding shows that even the highest mode of data privacy (random synthesis) is able to produce comparable results to less randomized sampling strategies and supports growing research about deriving meaningful conclusions in anonymized (or inaccurate) data.

3.7.1 References

Aiemagno, S. A. et al. 1996. "Assessing Substance Abuse Treatment Needs Among the Homeless: A Telephone-Based Interactive Voice Response System." *American Journal of Public Health* 86(11):1626–8. Retrieved May 11, 2017 (`http://ajph.aphapublications.org/doi/10.2105/AJPH.86.11.1626`).

Anderberg, Michael R. 1973. *Cluster Analysis for Applications*. Academic Press, Inc., New York.

Anderlucci, Laura and Christian Hennig. 2014. "The Clustering of Categorical Data: A Comparison of a Model-Based and a Distance-Based Approach." *Communications in Statistics-Theory and Methods* 43(4):704–21.

Anderson, Isobel. 2001. "Pathways Through Homelessness: Towards a Dynamic Analysis."

Arce, A. Anthony and Michael J. Vergare. 1984. "Identifying and Characterizing the Mentally Ill Among the Homeless." *The Homeless Mentally Ill: A Task Force Report of the American Psychiatric Association. Washington: American Psychiatric Association* 75–89.

Barrientos, Andrés F., Jerome P. Reiter, Ashwin Machanavajjhala, and Yan Chen. 2017. "Differentially Private Significance Tests for Regression Coefficients." *arXiv preprint arXiv:1705.09561*.

Bassuk, Ellen. 1984. "Is Homelessness a Mental Health Problem?" *American Journal of Psychiatry* 141(12):1546–50. Retrieved May 13, 2017 (`http://psychiatryonline.org/doi/abs/10.1176/ajp.141.12.1546`).

Bassuk, Ellen and Lenore Rubin. 1987. "Homeless Children: A Neglected Population." *American Journal of Orthopsychiatry* 57(2):279–86. Retrieved April 13, 2017 (`http://doi.apa.org/getdoi.cfm?doi=10.1111/j.1939-0025.1987.tb03538.x`).

Bassuk, Ellen L., Carmela J. DeCandia, Corey Anne Beach, and Fred Berman. 2014. "Amer-

ica's Youngest Outcasts: A Report Card on Child Homelessness."

Bassuk, Ellen, Katherine T Volk, and Jeffrey Olivet. 2010. "A Framework for Developing Supports and Services for Families Experiencing Homelessness." *The Open Health Services and Policy Journal* 3(1).

Benjaminsen, Lars and Stefan Bastholm Andrade. 2015. "Testing a Typology of Homelessness Across Welfare Regimes: Shelter Use in Denmark and the USA." *Housing Studies* 30(6):858–76. Retrieved May 13, 2017 (`http://www.tandfonline.com/doi/full/10.1080/02673037.2014.982517`).

Burt, Martha R. 2001. "Helping America's Homeless."

Burt, Martha R. and others. 2006. "Characteristics of Transitional Housing for Homeless Families." *Washington, DC: Urban Institute*.

Caton, Carol LM, Carol Wilkins, and Jacquelyn Anderson. 2007. "People Who Experience Long-Term Homelessness: Characteristics and Interventions." in *Toward understanding homelessness: The 2007 national symposium on homelessness research*. US Department of Health; Human Services; Department of Housing; Urban Development Washington, DC.

Chawla, Shuchi, Cynthia Dwork, Frank McSherry, and Kunal Talwar. 2005. "On Privacy-Preserving Histograms." in *Uncertainty in Artificial Intelligence (UAI)*. Edinburgh, Scotland: Association for Uncertainty in Artificial Intelligence. Retrieved (`https://www.microsoft.com/en-us/research/publication/on-privacy-preserving-histograms/`).

Culhane, Dennis P., Stephen Metraux, Jung Min Park, Maryanne Schretzman, and Jesse Valente. 2007. "Testing a Typology of Family Homelessness Based on Patterns of Public Shelter Utilization in Four U.S. Jurisdictions: Implications for Policy and Program Planning." *Housing Policy Debate* 18(1):1–28. Retrieved January 30, 2017 (`http:`

//www.tandfonline.com/doi/abs/10.1080/10511482.2007.9521591).

Culhane, Dennis P., Jung Min Park, and Stephen Metraux. 2011. "The Patterns and Costs of Services Use Among Homeless Families." *Journal of Community Psychology* 39(7):815–25. Retrieved May 8, 2017 (http://doi.wiley.com/10.1002/jcop.20473).

Danseco, Evangeline R. and E. Wayne Holden. 1998. "Are There Different Types of Homeless Families? A Typology of Homeless Families Based on Cluster Analysis." *Family Relations* 47(2):159. Retrieved May 8, 2017 (http://www.jstor.org/stable/585620?origin=crossref).

Dean, Nema and Adrian E. Raftery. 2010. "Latent Class Analysis Variable Selection." *Annals of the Institute of Statistical Mathematics* 62(1):11–35. Retrieved May 11, 2017 (http://link.springer.com/10.1007/s10463-009-0258-9).

Draine, Jeffrey, Mark S. Salzer, Dennis P. Culhane, and Trevor R. Hadley. 2002. "Role of Social Disadvantage in Crime, Joblessness, and Homelessness Among Persons with Serious Mental Illness." *Psychiatric Services* 53(5):565–73. Retrieved May 13, 2017 (http://psychiatryonline.org/doi/abs/10.1176/appi.ps.53.5.565).

Dwork, Cynthia. 2006. "Differential Privacy." Pp. 1–12 in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, vol. 4052. Venice, Italy: Springer Verlag. Retrieved (https://www.microsoft.com/en-us/research/publication/differential-privacy/).

Dwork, Cynthia. 2007. "Ask a Better Question, Get a Better Answer A New Approach to Private Data Analysis." Pp. 18–27 in *11th International Conference on Database Theory (ICDT 2007)*, vol. 4353. Barcelona, Spain: Springer. Retrieved (https://www.microsoft.com/en-us/research/publication/ask-a-better-question-get-a-better-answer-a-

Fazel, Seena, John R. Geddes, and Margot Kushel. 2014. "The Health of Homeless People in High-Income Countries: Descriptive Epidemiology, Health Consequences, and Clinical

and Policy Recommendations." *The Lancet* 384(9953):1529–40. Retrieved February 7, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S0140673614611326).

Fertig, Angela R. and David A. Reingold. 2008. "Homelessness Among at-Risk Families with Children in Twenty American Cities." *Social Service Review* 82(3):485–510. Retrieved May 13, 2017 (http://www.journals.uchicago.edu/doi/10.1086/592335).

Figueiredo, M.A.T. and A.K. Jain. 2002. "Unsupervised Learning of Finite Mixture Models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(3):381–96. Retrieved July 3, 2017 (http://ieeexplore.ieee.org/document/990138/).

Fischer, Pamela J. and William R. Breakey. 1991. "The Epidemiology of Alcohol, Drug, and Mental Disorders Among Homeless Persons." *American Psychologist* 46(11):1115–28. Retrieved April 5, 2017 (http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.46.11.1115).

Fraley, Chris and Adrian E. Raftery. 2006. *MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering.* University of Washington - Statistics Department.

Fraley, Chris, AE Raftery, and L. Scrucca. 2012. "Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." *Department of Statistics, University of Washington* 23:2012.

Galea, Sandro and David Vlahov. 2002. "Social Determinants and the Health of Drug Users: Socioeconomic Status, Homelessness, and Incarceration." *Public health reports* 117(Suppl 1):S135.

Gleason, Kristen, John P. Barile, and Charlene K. Baker. 2017. "Describing Trajectories of Homeless Service Use in Hawai'i Using Latent Class Growth Analysis." *American Journal of Community Psychology* 59(1-2):158–71. Retrieved May 12, 2017 (http://doi.wiley.

com/10.1002/ajcp.12128).

Grant, Roy, Delaney Gracy, Grifin Goldsmith, Alan Shapiro, and Irwin E. Redlener. 2013. "Twenty-Five Years of Child and Family Homelessness: Where Are We Now?" *American Journal of Public Health* 103(S2):e1–e10. Retrieved April 8, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.2013.301618).

Grigsby, Charles, Donald Baumann, Steven E. Gregorich, and Cynthia Roberts-Gray. 1990. "Disaffiliation to Entrenchment: A Model for Understanding Homelessness." *Journal of Social Issues* 46(4):141–56. Retrieved May 8, 2017 (http://doi.wiley.com/10.1111/j.1540-4560.1990.tb01803.x).

Hennig, Christian and Tim F. Liao. 2010. *Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type Variables with Application to Social Stratification.* Technical report.

Hertzberg, Edwina L. 1992. "The Homeless in the United States: Conditions, Typology and Interventions." *International Social Work* 35(2):149–61.

Humphreys, Keith. 1995. "Sequential Validation of Cluster Analytic Subtypes of Homeless Veterans." *American Journal of Community Psychology* 23(1):75–98. Retrieved May 8, 2017 (http://doi.wiley.com/10.1007/BF02506923).

Koh, H. K., G. Graham, and S. A. Glied. 2011. "Reducing Racial and Ethnic Disparities: The Action Plan from the Department of Health and Human Services." *Health Affairs* 30(10):1822–9. Retrieved June 20, 2017 (http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2011.0673).

Kuhn, Randall and Dennis P. Culhane. 1998. "Applying Cluster Analysis to Test a Typology of Homelessness by Pattern of Shelter Utilization: Results from the Analysis of Administrative Data." *American Journal of Community Psychology* 26(2):207–32. Retrieved

May 8, 2017 (http://doi.wiley.com/10.1023/A:1022176402357).

Lee, Barrett A., Kimberly A. Tyler, and James D. Wright. 2010. "The New Homelessness Revisited." *Annual Review of Sociology* 36(1):501–21. Retrieved April 8, 2017 (http://www.annualreviews.org/doi/10.1146/annurev-soc-070308-115940).

Link, B. G. et al. 1994. "Lifetime and Five-Year Prevalence of Homelessness in the United States." *American Journal of Public Health* 84(12):1907–12.

MacQueen, James and others. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." Pp. 281–97 in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Oakland, CA, USA.

Masten, A. S. et al. 2012. "Executive Function Skills and School Success in Young Children Experiencing Homelessness." *Educational Researcher* 41(9):375–84. Retrieved April 5, 2017 (http://edr.sagepub.com/cgi/doi/10.3102/0013189X12459883).

McLachlan, Geoffrey and David Peel. 2004. *Finite Mixture Models*. John Wiley & Sons.

Metraux, Stephen and Dennis P. Culhane. 1999. "Family Dynamics, Housing, and Recurring Homelessness Among Women in New York City Homeless Shelters." *Journal of family issues* 20(3):371–96.

Metraux, Stephen and Yi-Ping Tseng. 2017. "Using Administrative Data for Research on Homelessness: Applying a US Framework to Australia."

Miller, P. M. 2011. "A Critical Analysis of the Research on Student Homelessness." *Review of Educational Research* 81(3):308–37. Retrieved May 8, 2017 (http://rer.sagepub.com/cgi/doi/10.3102/0034654311415120).

Morgenstern, Jon et al. 2006. "Effectiveness of Intensive Case Management for Substance-Dependent Women Receiving Temporary Assistance for Needy Families." *American Journal of Public Health* 96(11):2016–23. Retrieved June 20, 2017 (http://ajph.aphapublications.

org/doi/10.2105/AJPH.2005.076380).

Morrison, D. S. 2009. "Homelessness as an Independent Risk Factor for Mortality: Results from a Retrospective Cohort Study." *International Journal of Epidemiology* 38(3):877–83. Retrieved April 5, 2017 (https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyp160).

Morse, Gary A. 1992. "Causes of Homelessness." Pp. 3–17 in *Homelessness: A National Perspective*, edited by M. J. Robertson and M. Greenblatt. Boston, MA: Springer US. Retrieved (http://dx.doi.org/10.1007/978-1-4899-0679-3_1).

Morse, Gary A., Robert J. Calsyn, and Gary K. Burger. 1992. "Development and Cross-Validation of a System for Classifying Homeless Persons." *Journal of Community Psychology* 20(3):228–42. Retrieved (http://dx.doi.org/10.1002/1520-6629(199207)20:3<228::AID-JCOP2290200306>3.0.CO;2-0).

Mowbray, Carol T., Deborah Bybee, and Evan Cohen. 1993. "Describing the Homeless Mentally Ill: Cluster Analysis Results." *American Journal of Community Psychology* 21(1):67–93. Retrieved May 8, 2017 (http://doi.wiley.com/10.1007/BF00938208).

Nunez, Ralph and Cybelle Fox. 1999. "A Snapshot of Family Homelessness Across America." *Political Science Quarterly* 114(2):289–307. Retrieved April 8, 2017 (http://doi.wiley.com/10.2307/2657740).

O'Connell, James J. 2005. "Premature Mortality in Homeless Populations: A Review of the Literature." *Nashville, TN: National Health Care for the Homeless Council*.

O'Connell, Mary Ellen. 2003. "Responding to Homelessness: An Overview of US and UK Policy Interventions." *Journal of Community & Applied Social Psychology* 13(2):158–70. Retrieved June 20, 2017 (http://doi.wiley.com/10.1002/casp.720).

Page, Timothy and Roger M. Nooe. 2002. "Life Experiences and Vulnerabilities of Homeless Women: A Comparison of Women Unaccompanied Versus Accompanied by Minor

Children, and Correlates with Children's Emotional Distress." *Journal of Social Distress and the Homeless* 11(3):215–31.

Ping, Haoyue, Julia Stoyanovich, and Bill Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." P. 42 in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Proust-Lima, Cécile, Viviane Philipps, and Benoit Liquet. 2015. "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package Lcmm." *arXiv preprint arXiv:1503.00890*.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved (`https://www.R-project.org/`).

Rafferty, Yvonne and Marybeth Shinn. 1991. "The Impact of Homelessness on Children." *American Psychologist* 46(11):1170–9.

Rasmussen, Carl Edward. 2000. "The Infinite Gaussian Mixture Model." Pp. 554–60 in *Advances in neural information processing systems*.

Reiter, Jerome P. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18(4):531.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of official Statistics* 9(2):461–68.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8(1):289.

Shinn, Marybeth et al. 2008. "Long-Term Associations of Homelessness with Children's Well-Being." *American Behavioral Scientist* 51(6):789–809. Retrieved April 8, 2017

(http://journals.sagepub.com/doi/10.1177/0002764207311988).

Shlay, Anne B. and Peter H. Rossi. 1992. "Social Science Research and Contemporary Studies of Homelessness." *Annual Review of Sociology* 18(1):129–60. Retrieved May 9, 2017 (http://www.annualreviews.org/doi/10.1146/annurev.so.18.080192.001021).

Solarz, Andrea and G. Anne Bogat. 1990. "When Social Support Fails: The Homeless." *Journal of Community Psychology* 18(1):79–96. Retrieved (http://dx.doi.org/10.1002/1520-6629(199001)18:1<79::AID-JCOP2290180112>3.0.CO;2-B).

Wood, David L., R. Burciaga Valdez, Toshi Hayashi, and Albert Shen. 1990. "Health of Homeless Children and Housed, Poor Children." *Pediatrics* 86(6):858–66.

Chapter 4

FAMILY DEMOGRAPHIC CHARACTERISTICS AND SUCCESSFUL PATHWAYS TO EXITING HOMELESSNESS PRE- AND POST-GREAT RECESSION

4.1 Introduction

Policy makers and researchers have identified housing assistance programs as integral support services that create pathways for individuals and families to exit homelessness. The Departments of Housing and Urban Development and Health and Human Services sponsor numerous assistance and surveillance programs such as rapid rehousing and transitional housing to diminish homelessness duration and episodes. While macro-forces such as economic conditions (income inequality, affordable housing supply, shifting economy, etc.), insufficient social welfare programs, and incarceration are principle drivers of homelessness, micro-level factors such as individual health (e.g., mental health disorders), family characteristics (household structure, race, employment, etc.), and homeless assistance programs also influence homelessness duration and likelihood of exiting homeless cycles (Aiemagno et al. 1996; Bassuk 1984; Draine et al. 2002; Galea and Vlahov 2002; Gould and Williams 2010; Hertzberg 1992; Page and Nooe 2002; Toro and Wall 1991).

Multiple disputes in homelessness research create opportunities for novel data and methods to potentially understand homelessness cyclicity by analyzing exit pathways associated with homelessness interventions. For instance, researchers have disputed how influential housing program and family characteristics are for homelessness duration or persistence in homeless cycles making targeted interventions difficult. Typically, three types of housing-first programs (rapid re-housing, transitional housing, and emergency

housing) are used to service homeless families. There is a strong belief in homelessness research that housing-first programs are best (Lee, Tyler, and Wright 2010); however, program efficacy across different household characteristics is not clear (Grant et al. 2013). Understanding the relationship between housing-first programs and exits from homeless cycles (successful exit, successful exit with subsidy, and other exit) can influence targeted service delivery and public investment in homelessness resources. Furthermore, the Great Recession (2007-2009) has changed the composition, health, and use of social services amongst homeless individuals and families in ways that are not fully understood (Bennett, Scharoun-Lee, and Tucker-Seeley 2009; Pilkauskas, Currie, and Garfinkel 2012; Sard 2009; Treas 2010). Recent economic distress has changed the composition of homeless populations and the potential post-Great Recession demographic shifts may have impacted homelessness service use (Miller 2011).

Lastly, homeless populations are difficult to track and improved data/methods are needed to evaluate interventions and supplement housing assistance policy. Link et al. (1994) describes multiple concerns that plague homelessness surveys and research such as finding currently homeless individuals and reluctance to participate. Additionally, homeless individuals may not be completely forthcoming with providers while registering for services and intentionally present inaccurate demographic and lifestyle information (e.g., substance use) (Aiemagno et al. 1996; Metraux, Stino, and Culhane 2014; Midderhoff 1994). These data limitations suggest that homeless data analyses should account for demographic background uncertainty while evaluating homelessness program efficacy. Advancements in creating synthetic data have shown that these data can maintain client privacy for individuals using homelessness services and introduce data uncertainty (noise) that is comparable to challenges introduced by variable data quality and respondent accuracy.

To address gaps in family homeless research, this study compares housing assistance programs and investigates the role of family characteristics on exit pathways. Using

synthetic Homelessness Management Information Systems (henceforth HMIS) data, the joint influence of family demographic characteristics and housing-first program type on homelessness exit pathways are assessed. Synthetic HMIS data will help provide client privacy and account for family demographic uncertainty with differential privacy (noise added to underlying data). Random forest classifiers predict how combinations of housing-first programs will affect different families' likelihood of exiting homelessness cycles. These analyses reveal the chances of successful exit vary by program type and family background. Tremendous (time and geographic) uncertainty regarding program administration makes it difficult to pinpoint causes of varying success rates. However, the empirical findings of inconsistent outcomes across demographic groups informs the need to tailor housing programs and case management to family backgrounds.

4.2 Background

This review discusses the state-funded landscape for homelessness assistance programs, the need to identify effective solutions, and the potential effects of the Great Recession on homeless family resource utilization. A consistent working definition of homelessness is among the many reasons homelessness intervention programs face difficulty. Defining homelessness is important because program enrollment varies based on eligibility. The two agencies primarily tasked with homeless assistance, the Department of Housing and Urban Development (henceforth HUD) and Department of Health and Human Services (henceforth HHS), utilize overlapping but inconsistent definitions of homelessness. While HHS favors an expansive definition of homelessness that acknowledges precarious housing arrangements such as doubling up[1], HUD favors a narrow definition consistent with the structurally unhoused. Not surprisingly, quantitative methods to identify and assess homeless populations are encumbered by measurement problems (Anderson 2003).

[1] "doubling up refers to a situation where individuals are unable to maintain their housing situation and are forced to stay with a series of friends and/or extended family members" (`https://www.nhchc.org/faq/official-definition-homelessness/`)

Additional research is needed to better understand the role of housing and services in stabilizing different subgroups of families, as well as which approaches are most cost-effective (Bassuk and Geller 2006).

4.2.1 Homelessness Assistance Programs

Modern funding for homelessness programs dates back to the 1987 McKinney-Vento Act designed to increase funding to homeless shelters and provide for the educational needs of homeless children. To supplement the McKinney-Vento Act, HUD began advocating for continuum of care programs that build a pipeline between housing aid and homelessness services (HUD 2002). McKinney-Vento originally consisted of fifteen program types but has been trimmed to focus on Homeless Management Information Systems (HMIS) for data infrastructure, supportive services, and three housing prevention/intervention programs: permanent supportive housing, rapid rehousing, and transitional/emergency housing[2]. In recent legislation such as the 2009 Homeless Emergency Assistance and Rapid Transition to Housing (HEARTH) Act and internal policy, HUD has favored assistance programs that use permanent housing support (HUD Funding notice 2015). Permanent housing is defined by HUD as "community-based housing without a designated length of stay in which formerly homeless individuals and families live as independently as possible"[3]. Continuum of care approaches have also lead to collaboration between child welfare and public housing services (Fowler and Chavira 2014). Assessing the efficacy of housing intervention programs as well as HMIS data infrastructure can improve the delivery of homeless services.

---

[2]`https://www.hudexchange.info/programs/coc/coc-program-eligibility-requirements/`. Permanent supportive housing is sometimes interchanged with permanent housing throughout the text.

[3]`https://www.hudexchange.info/programs/coc/coc-program-eligibility-requirements/`

4.2.2 Homeless Management Information Systems

HUD funding eligibility requires homeless service providers meet data reporting standards about homeless individuals served. HMIS is an electronic record system that meets HUD data reporting standards and is designed for case management (locally, within an organization) and high-level views of homeless populations (globally, when aggregated by various levels of government such as city, county, state, etc.). Regional HMIS are faced with technological and human resource disadvantages that impact the continuum of care for homeless individuals (Zhang and Gutierrez 2007). For instance, regional HMIS are tasked with understanding a complex web of hundreds of homeless assistance programs, handling confidential data, and screening for unreliable data entry by both individuals and service provider. These responsibilities make tracking pathways to existing homelessness difficult (Cronley and Patterson 2012). Moreover, the allocation of individuals and families to homeless assistance programs varies by demographic characteristics such as age, race, mental health, and veteran status (Montgomery et al. 2016; Washington et al. 2010). Although families are not randomly assigned to housing programs, natural variation in program population allows for comparisons between programs and across multiple family types.

4.2.3 Housing Assistance Program Debate

The three main housing-first assistance programs are permanent, rapid, and transitional rehousing. The National Healthcare for the Homeless Council defines permanent supportive housing as "a model that combines low-barrier affordable housing, health care, and supportive services to help individuals and families lead more stable lives"[4]. Rapid rehousing and transitional rehousing[5] are usually time-limited interventions to provide housing services for chronic and acute housing dilemmas such as crisis relief for groups like working families struggling to make rent and programs for individuals addressing

[4] https://www.nhchc.org/policy-advocacy/issue/permanent-supportive-housing/

[5] http://www.transitionalhousing.org/

barriers to housing (e.g., substance abuse, mental health concerns) (USICH 2014). Emergency shelters are extremely short-term housing interventions such as overnight shelters with more limited support services and case management than the three housing-first interventions. These housing interventions are designed to address the various needs of homeless families from the transitional, episodic, and chronically homeless. Transitionally homeless families typically experience a single, short episode of homelessness, episodically homeless families have few episodes of short stays, and chronically homeless families have multiple episodes of homelessness of lengthy duration (Culhane et al. 2007).

Housing-first program efficacy and the role of family background is debated in homelessness literature. For instance, Culhane and Metraux (2008) find that emergency and transitional services are not adequately addressing homeless needs and Fisher et al. (2014) reports that "respondents were least comfortable in and most likely to leave transitional housing"(pg. 1). Researchers are concerned that some assistance programs do not positively affect homeless experiences such as reducing the length of shelter stay and decreasing the probability of homelessness resource use in the future (Goodman, Messeri, and O'Flaherty 2014). Comparing housing-first programs' probability of successful exits across demographic and geographic groups can help clarify housing-first program efficacy. Along, with debates over housing program efficacy, homelessness research is concerned with the ability to match services to individual or family background.

4.2.4 Matched Services

Matching intervention programs and participants improves outcomes for individuals and organizations (Choi and Ryan 2007). Program matching refers to the practice of targeted program placement for individuals or families based on demographic characteristics such as race, gender, household structure, employment status, etc. These matching approaches were undertaken because previous data demonstrated strong relationships between individual/family characteristics and program success. Moreover, an additional result of well matched programs is the successful reunification of vulnerable families separated by mul-

tiple interventions.

Culhane et al. (2007) argues that "program and policy factors appear to play a primary role in shaping shelter utilization" (pg. 26). Shinn (1997) also claims that the "[r]eceipt of subsidized housing, in turn, was predicted primarily by the shelter to which families were assigned and their length of stay, factors that were unrelated to family characteristics" (pg. 760). Conversely, some researchers claim that individual and family characteristics matter more for exiting homelessness cycles into stable housing than assistance program used. Metraux and Culhane (1999) is emblematic of literature that contends family dynamics such having young children, larger families, older head of household, domestic violence, etc. (these authors, like many in this area, believe the availability of affordable housing is the largest factor driving shelter usage). Wood et al. (1990) and Rossi (1991) offer additional support for family dynamics influencing shelter utilization patterns. Homeless families are generally headed by women which has increased the rates of children experiencing homelessness (Bassuk and Rosenberg 1988; Bassuk et al. 1997, 2014; Fertig and Reingold 2008; Nunez and Fox 1999). Given that family demographic differences (gender, race of parents, size of family, etc.) are influential for finding stable housing, supportive programs for single female headed households is a pressing concern (Bassuk et al. 1997, 2006; Rocha et al. 1996; Walsh et al. 2014; Wong, Culhane, and Kuhn 1997).

4.2.5 Great Recession

Further exploring the resource utilization of families pre- and post-Great Recession is worthwhile to understand the compositional changes to homeless families and potential differences in family resource service use. Rising foreclosure and unemployment during the global Great Recession (2007-2009) affected the composition of homeless individuals and households (Elsby, Hobijn, and Sahin 2010). Available research on compositional changes to homeless individuals and families suggest limited affordable housing in conjunction with diminished employment opportunities dramatically increased shares of homeless in-

dividuals and families (Bassuk et al. 2014; Ellen and Dastrup 2012; Miller 2011; Oberg 2011). Ellen and Dastrup (2012) cite a 30% rise in homeless families from 2007 to 2009 alone with 40,000 families becoming homeless and thousands more families in precarious housing situations.

4.3 Data and Methods

HMIS data were collected from multiple service providers in HUD and HHS funded programs such as transitional housing, emergency shelters, rapid-rehousing, support services only, etc. Each observation represents a unique family's homeless experience (duration, programs used), and respondent's living situation at the time of program exit. The limitations of administrative data can be directly addressed by methodological advances in producing synthetic data and clustering. Rubin (1993) proposed creating synthetic data sets to address confidentiality concerns for secure data. Synthetic data sets are designed to ignore sampling schema, use multiple imputation, and most importantly for this study incorporate intentionally inaccurate (e.g., not the actual collected value in the original data set) data for some variables. Reiter (2002) and others have found that valid inferences can be gained from synthetic data. Given the concerns with data and respondent reliability in HMIS, using intentionally inaccurate data to model relationships builds robustness. Models that project forward with HMIS data should be sensitive to unknown, inaccurate data instead of face-value acceptance of the HMIS data quality. Comparing programs across multiple synthetic data types can help build confidence that an intervention is preferable to other alternatives.

Synthetic data is used to maintain the privacy of individuals in the HMIS data and introduce data uncertainty that mimics real word data entry and respondent accuracy limitations. A major challenge with synthetic data is that realistic enough data potentially violate individuals' privacy (for example, by saying things about their neighborhood). On the other hand, if the data is obfuscated, it might become less useful for drawing any conclusions. Cryptographic research has developed methods such as differential privacy

which demonstrate the ability to preserve client privacy, while maintaining a connection to the underlying data (Chawla et al. 2005; Dwork 2006, 2007). Given difficulties with retroactive use of data (data entry, data quality, etc.) cryptographic and machine learning methods that are novel to homelessness research could clarify inferences (Metraux and Tseng 2017). Furthermore, algorithmic bias with sensitive data requires HMIS funded organizations adopt a unique workflow because these HMIS programs cannot give analysts private data to develop tools. Instead, homelessness service providers can leverage differential privacy by masking the "true" value of any respondent covariate with noise and missingness and still allow analysts to develop tools that can then be applied to real data (Barrientos et al. 2017). Respect for data sharing agreements (data sensitivity) along real-world limitations with data entry (data accuracy) make HMIS data ideal for synthetic analysis.

The synthetic data sets were created by using the open source software DataSynthesizer[6]. DataSynthesizer can synthesize data in three different modes that add varying degrees of differential privacy (noise) (Ping, Stoyanovich, and Howe 2017). The basic approach for each mode is the same (create a distribution for each variable and samples that distribution); however, the degree of differential privacy and missing rates used to further anonymize the data change. The software for creating synthetic data has a noise parameter that is equivalent to the minimum change in the correlations by removing (perturbing) any tuple of data at random. Missingness is determined by the synthetic mode utilized. In the random mode, ideal for the most sensitive data, the DataSynthesizer generates type consistent random variables, adding the highest degree of differential privacy. The correlated attribute mode uses a Bayesian network to calculate the relationship between variables and in cases where calculating a Bayesian network is too computationally expensive Ping et al. (2017) suggest using the independent attribute to sample from a noise-added distribution of the underlying data. After synthesizing a single county's HMIS

[6]`https://github.com/DataResponsibly/DataSynthesizer`

(and sampling an untold percentage of this county)[7], the synthesized data was synthesized again. Thus, the final data set is a twice synthesized representation of HMIS data and is a definitive abstraction of any county HMIS data. The synthesized independent data and synthesized random data are similar in size (these synthesized datasets are 120% the size of synthesized correlated data) and these jurisdictions are based on longitudinal program data from 1993 to 2015.

4.3.1 Analytical strategy

The focal outcome use the client's post-program plans (destinations or exits) which are reported during an exit interview, if the exit interview takes place. Some clients utilized multiple homeless services and the last provider type is associated with the exit. A successful exit is defined as: 'Rental by client, no ongoing housing subsidy', and 'Staying or living with friends, permanent tenure'. Successful exit with a subsidy includes destinations such as: 'Rental by client, with other ongoing housing subsidy', 'Permanent housing for formerly homeless persons and 'Rental by client, with Veteran Affairs Supportive Housing (VASH) housing subsidy'. Lastly other exits were: 'Emergency shelter, including hotel or motel paid for with emergency shelter voucher', 'Hotel or motel paid for without emergency shelter voucher', 'Hospital or other residential non-psychiatric medical facility', 'Substance abuse treatment facility or detox center', 'Staying or living with family, temporary tenure (e.g., room, apartment or house)' as well as data collection limitations (e.g., 'no exit interview completed', 'client refused', and 'data not collected').

Random forests are an ensemble machine learning method that use decision trees for (categorical and continuous) prediction tasks (Breiman 2001). Classification decision trees mine data to predict a categorical label that corresponds with input variables while using randomization and data cross-validation (training and test sets) to account for the decision tree's tendency to overfit models (Breiman 1996). Decision trees randomly

[7]Due to the sensitivity of the data, the amount sampled will not be revealed.

select observations with a criteria (e.g., random selection without replacement)[8] from the training set to build a predictive model and the strength of the predictive relationship is assessed in the test set of data that was not used to train the predictive model (the training data). The ensemble aspect of random forests describes the process of using the mode (classification) or mean (regression) of multiple decision trees to select the best model for the data (Breiman 2001). Using the Python package `scikit-learn`, random forests classifiers predict the exit (successful, successful with subsidy, other) linked to the last housing assistance program for a family. The analysis was completed Python software (Rossum 1995) and relied heavily on the `sckit-learn` package for decision tree analysis (Pedregosa et al. 2011).

Defining the Great Recession

The data was split on the year 2007 to construct the pre- and post-Great Recession periods. All observations on or before 2007 comprise the pre-Great recession period. Typically, in the synthetic data, the minority of observations come from pre-Great Recession families (e.g., 5% of the entire data) while the post-Great Recession period includes all observations in 2008 or later (often around 95% of the data).

Exclusionary criteria

Demographic combinations (household structure and race) with less than five observations were removed because random forest out-of-sample cross validation methods require a minimum of five observations. Additionally, household combinations with unknown race were not reported. The reduced sample also only considers households with one or two adults (the 9 households with greater than two adults are likely data entry errors) as well as observations without any missingness on input and outcome variables.

---

[8]For reproducible analysis, the decision trees used the same random state instead of a randomly selecting new observations and training different classification models. All code and data will be available at `https://github.com/kpolimis`.

4.4 Findings

4.4.1 Overview

The synthetic datasets had varying observation due to way missingness is handled in each synthetic form. Synthetic data created in the random and independent attribute mode have the same amount of observations (4647) while the correlated synthetic data had the fewest observations (2726). In the random synthetic data, 67.7% of the data is pre-Great Recession and 32.3% is post-Great Recession. Contrastingly, in the independent synthetic data, most of the data comes from the post-Great Recession period (86.46%) while 13.54% of the data is pre-Great Recession. Lastly, the correlated synthetic data is even more skewed towards the post-Great Recession period. In the correlated synthetic data, 99.12% of the observations are post-Great Recession and 0.88% of the data is pre-Great Recession (See Appendix C for all summary statistics tables and exit probability tables).

4.4.2 Synthetic Independent Attribute

pre-Great Recession

Overall, in pre-Great Recession period (Table 4.1), the predicted probability of successful exit was 0.45, successful exit with subsidy was 0.17 and other exit were 0.38 (All Tables are available in Appendix C). The program associated with the best probability of successful exit was emergency shelter housing (0.62). Emergency shelters were the only program with a probability of success greater than 50%. Permanent supportive housing had the best chance of successful exit with subsidy (0.3) while rapid rehousing (0.52) and permanent supportive housing were the housing programs that had the highest probability of other exits (0.45).

Table 4.1: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.52	0.41	0.07
Transitional Housing	0.31	0.49	0.2
Emergency Shelter	0.25	0.62	0.13
Permanent Housing	0.45	0.26	0.3

Single female households were consistent with the overall trends and emergency shelter housing was associated with the largest probability of successful (0.45) and successful exit with subsidy (0.4). Rapid rehousing had the highest probability of other exit (0.57). However, four of the six racial groups associated a program besides emergency shelters as the program best linked with exiting homelessness. Single black and multi-racial females are predicted to successfully exit homeless 100% of the time in emergency shelter housing and these observations skew the single female findings overall.

Amongst coupled households, emergency housing was associated with the largest probability of successful exit (0.83) while transitional housing had the highest probability of successful with subsidy (0.2). Similar to the overall pre-Great Recession trend, rapid rehousing and permanent supportive housing were associated with the greatest probability of other exit (0.5 and 0.47; respectively). Similar to single females, racial variation was observed within this household type and some racial-household combinations were more successful exiting homeless in intervention programs other than emergency shelters. Overall, across all household types, emergency corresponded with the greatest predicted probability of successful exit, transitional housing had the best chances of successful exit with subsidy, and rapid rehousing and permanent supportive housing were most associated with other exits.

post-Great Recession

Overall, in the post-Great Recession period (Table 4.2), the predicted probability of successful exit were 0.32, successful exit with subsidy was 0.17 and other exit were 0.5. Homeless families in the post-Great recession period had lower probabilities of successfully exiting homelessness, equal chances of successful exit with subsidy, and greater probabilities of other exit when compared to the pre-Great Recession period.

The programs associated with the best probability of successful exit were rapid rehousing (0.4), followed closely by transitional housing (0.39). Participation in rapid rehousing (0.22) and transitional housing (0.22) had the best chance of successful exit with subsidy, emergency shelters were the housing program that had the highest probability of other exits (0.69).

Table 4.2: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.4	0.19
Transitional Housing	0.4	0.39	0.22
Emergency Shelter	0.69	0.19	0.11
Permanent Housing	0.5	0.32	0.18

Single female households participating in rapid rehousing were associated with the largest probability of successful exit (0.37). Emergency shelters were associated with the largest probability of successful exit with subsidy (0.22) and the highest probability of other exit (0.56). For single male households, transitional housing was associated with the largest probability of successful exit (0.56) and the largest probability of successful exit with subsidy (0.28). Emergency shelter had the highest probability of other exit (0.88). While all housing interventions had a near equal probability of other exit for single females,

emergency shelters were demonstrably worse than any other intervention for single males. Most of the racial groups within single female and male households aligned with the household trend. Amongst coupled households, rapid rehousing was associated with the highest probability of a successful exit (0.37) and successful exit with subsidy (0.22) while emergency shelter had the highest probability of other exit (0.63). Racial comparisons revealed that couples had more variation for housing program associated with highest probability of successful exit than single households. Across two of the three household types, rapid rehousing corresponded with the greatest predicted probability of successful exit and numerous programs vied for most likely to produce a successful exit with subsidy. Emergency shelters were strongly associated with other exits for two of the three household types.

Entire data

Lastly, for the entire data (Table 4.3), the predicted probability of successful exit were 0.33, successful exit with subsidy was 0.17 and other exit were 0.5. The program associated with the best probability of success was rapid rehousing (0.42) closely followed by transitional housing (0.39). Participation in transitional housing and permanent supportive housing had near equal chance of successful exit with subsidy (0.21 and 0.19, respectively). Emergency shelters were the housing program that had the highest probability of other exits (0.67).

Table 4.3: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.42	0.16
Transitional Housing	0.39	0.39	0.21
Emergency Shelter	0.67	0.21	0.12
Permanent Housing	0.51	0.31	0.19

The post-Great Recession group are the largest percentage of observations and drive the trends seen in the entire data. Single female households participating in rapid rehousing were associated with the largest probability of any successful exit (0.51). Emergency shelters had the highest probability of other exit (0.54). For single male households, transitional housing was associated with the largest probability of any successful exit (0.6). Emergency shelter had by far the highest probability of other exit (0.88). Amongst coupled households, rapid rehousing was associated with the largest probability of any successful exit (0.63). Emergency shelter programs had the greatest probabilities of other exit (0.59). The household variations mirrored the findings from the post-Great Recession period.

4.4.3 Synthetic Correlated Attribute

pre-Great Recession

Overall, in pre-Great Recession period (Table 4.4), the predicted probability of successful exit was 0.22, successful exit with subsidy was 0.3 and other exit were 0.49 (All Tables are available in Appendix C). The program associated with the best probability of a successful exit was emergency shelters (0.55) by a wide margin (0.37) Participation in transitional housing had the best chance of successful exit with subsidy (0.5), permanent supportive housing lead all other programs in the probability of other exits (0.81) by a wide margin as well (0.41).

Table 4.4: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.38	0.18	0.44
Transitional Housing	0.4	0.1	0.5
Emergency Shelter	0.35	0.55	0.1
Permanent Housing	0.81	0.04	0.15

Single female households in emergency shelters were associated with the largest probability of successful exit (0.55). Transitional housing was associated with the largest probability of successful exit with subsidy (0.5) and the highest probability of other exit (0.81). Only one of the five racially distinct households did not achieve the highest probability of successful exit with emergency shelters. Lastly, other household types did not have enough observations to estimate predictions.

post-Great Recession

Overall, post-Great Recession period (Table 4.5), the predicted probability of successful exit was 0.36, successful exit with subsidy was 0.24 and other exit was 0.41. The program associated with the best probability of success was permanent supportive housing (0.41) followed closely by emergency shelters (0.38). Participation in transitional housing (0.32) had the best chance of successful exit with subsidy. Emergency shelters and rapid rehousing were the housing programs that had the highest probability of other exits (both 0.45).

Table 4.5: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.45	0.32	0.24
Transitional Housing	0.36	0.32	0.32
Emergency Shelter	0.45	0.38	0.18
Permanent Housing	0.38	0.41	0.21

Single female households participating in rapid rehousing were associated with the largest probability of a successful outcome (successful exit with and without subsidy) (0.38). Rapid rehousing was associated with the largest probability of successful exit with subsidy (0.45) and multiple programs had near equal probabilities of successful exit.

Emergency shelters had the highest probability of other exit (0.51). For single male households, permanent supportive housing was associated with the largest probability of a successful exit (0.44) and successful exit with subsidy (0.23). Rapid rehousing had the highest probability of other exit (0.51) although multiple programs had near 50% chance of other exit. Amongst coupled households, transitional rehousing was associated with the highest probability of any successful outcome (0.37 probability of successful exit with and without subsidy). Emergency shelters were associated with the largest probability of a successful exit (0.52) and transitional housing was associated with the largest probability of success with subsidy (0.61). Rapid rehousing had the highest probability of other exit (0.63) by a comfortable margin. Once again, the households analyzed exhibited racial variations in success probabilities. These findings are part of the theme across data types that show housing program success varies by household type.

Entire data

Lastly, for the entire data (Table 4.6), the predicted probability of successful exit were 0.36, successful exit with subsidy was 0.24 and other exit were 0.4. All housing programs were very similar in their probabilities for other, successful, and successful exit with subsidy. The program associated with the best probability of success was permanent supportive housing (0.41). With the synthetic correlated data, families in the pre-Great Recession period were less likely to successful exit homeless cycles (0.51) than families in the post-Great Recession era (0.59). The probabilities of successfully exiting homelessness were greater in the pre-Great Recession period for the synthetic independent data. Participation in transitional housing had the greatest chance of successful exit with subsidy (0.32) while the other housing programs had similar probabilities of successful exit with subsidy. Emergency shelters and rapid rehousing were the housing programs with the highest probability of other exits (0.45 and 0.42; respectively).

Table 4.6: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.36	0.21
Transitional Housing	0.36	0.32	0.32
Emergency Shelter	0.45	0.35	0.2
Permanent Housing	0.38	0.41	0.21

Single female households participating in rapid rehousing were associated with the largest probability of successful exit (0.46) and successful exit with subsidy (0.39). Emergency shelters had the highest probability of other exit (0.51). For single male households, permanent supportive housing had the greatest probability of successful exit (0.44 and successful exit with subsidy (0.23). Rapid rehousing and emergency shelters had similar probabilities of other exit (0.5 and 0.48; respectively). Amongst coupled households, transitional housing had the highest probability of any successful exit (0.38). However, emergency shelters and permanent supportive housing had high probabilities of successful exit (0.45 and 0.4). Rapid rehousing was associated with the largest probability of other exit for couples (0.62). Across single male and couples households in the correlated synthetic data, rapid rehousing was most associated with other exits and transitional housing was the best program predicting successful exits. This pattern did not hold for single female headed households.

4.4.4 Synthetic Random Attribute

pre-Great Recession

Overall, in pre-Great Recession period (Table 4.7), the predicted probability of successful exit was 0.14, successful exit with subsidy was 0.16 and other exit were 0.7 (Program Success Tables are available in Appendix C). Transitional housing had the best

chance of successful exit (0.16) and successful exit with subsidy (0.17). All housing programs had high probabilities of other exit (much higher than other modes of synthetic data), and rapid rehousing (0.74) had the highest probability of other exits.

Table 4.7: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.74	0.1	0.16
Transitional Housing	0.67	0.16	0.17
Emergency Shelter	0.7	0.14	0.17
Permanent Housing	0.7	0.15	0.15

Nearly all single females participating in rapid rehousing had other exits (0.9). Rapid rehousing was not an outlier, other housing programs had probabilities of other exit that exceeded 0.72. Single female households participating in emergency shelters were associated with the largest probability of any successful exit (0.28) and the other housing programs had near equal probabilities of successful exit with subsidy (exception, rapid rehousing). Single men also had high probabilities of other exit regardless of program type (0.7), though not as high as single women (0.78). Single men with children participating in transitional housing were associated with the greatest probability of any successful exit (0.37) although emergency shelters had the best probability of successful exit alone (0.18). Amongst coupled households, permanent supportive housing was associated with the largest probability of any successful exit (0.44) while emergency shelters had the highest probability of successful exit with subsidy (0.3). Depending on the household type in the random synthesized data, a different housing program was associated with the probability of successful exit in the pre-Great Recession period.

post-Great Recession

Overall, post-Great Recession period (Table 4.8), the predicted probability of successful exit were 0.18, successful exit with subsidy was 0.2 and other exit were 0.62. The probability of successful exit with and without subsidy increased from the pre- to post-Great Recession time periods a trend similar to the synthetic correlated data (but not the synthetic independent data). The program associated with the best probability of success was rapid rehousing (0.24), followed closely by transitional housing (0.12). Participation in rapid rehousing (0.28) and transitional housing (0.28) had the best chance of successful exit with subsidy, emergency shelters were the housing program that had the highest probability of other exits (0.65).

Table 4.8: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.65	0.24	0.11
Transitional Housing	0.6	0.12	0.28
Emergency Shelter	0.57	0.19	0.23
Permanent Housing	0.65	0.17	0.17

Single female households participating in emergency shelters were associated with the largest probability of successful exit (0.2) and successful exit with subsidy (0.34). Clients in rapid rehousing had the highest probability of other exit (1). For single male households, rapid rehousing was associated with the largest probability of successful exit (0.42) while emergency shelters had the largest probability of successful exit with subsidy (0.21). Permanent supportive housing had the highest probability of other exit (0.81). Coupled households enjoyed more successful exits (0.51) than either single females (0.29) or single males (0.34) Amongst coupled households, rapid rehousing was associated with the highest probability of a successful outcome (successful exit with and without subsidy) overall (0.5) and transitional housing had the highest probability of successful exit with

subsidy (0.42). Emergency shelters were associated with the largest probability of other exit (0.58). Each post-Great Recession households favored a different housing program when predicting the greatest probability of successful exit.

Entire data

Lastly, for the entire data in random synthetic mode (Table 4.9), the predicted probability of successful exit were 0.15, successful exit with subsidy was 0.18 and other exit were 0.67. All programs were near equal in probabilities of successful exit with permanent supportive housing the leading program (0.16). Participation in transitional housing and emergency shelters produced roughly equal chance of successful exit with subsidy (0.19 and 0.2; respectively). Rapid rehousing was the housing program that had the highest probability of other exits (0.71).

Table 4.9: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.71	0.14	0.15
Transitional Housing	0.67	0.14	0.19
Emergency Shelter	0.64	0.15	0.2
Permanent Housing	0.67	0.16	0.17

Single female households participating in rapid rehousing were associated with the largest probability of other exit (0.93). Contrastingly, emergency shelter programs had the largest probabilities of successful exit (0.15) and successful exit with subsidy (0.25). Unlike single female headed households, single male households in rapid rehousing programs had the greatest probability of successful exit (0.22 and transitional housing was associated with the largest probability of successful exit with subsidy (0.22). For single males, permanent supportive housing had the highest probability of other exit (0.78). Amongst

coupled households, permanent supportive housing was associated with the largest probability of any successful exit (0.43). Emergency shelters as a program had the highest probability of successful exit without subsidy (0.24). Transitional housing and emergency shelter programs had near equal probabilities of other exit (0.61 and 0.6). Household types in the synthesized random data were predicted to successfully exit homelessness cycles by enrolling in different housing programs. Permanent supportive corresponded with the greatest predicted probability of successful exit for single males and couples, transitional housing performed well for chances of successful exit with subsidy. Emergency shelters were surprisingly the most successful housing intervention for single female households.

In two of the three synthetic modes of data, homeless families had greater probabilities of successfully exit homelessness in the post-Great Recession period. However, the programs associated with successful exits varied by household type in each period and sometimes by racial group within household types. Demographically, post-Great Recession households have more single female heads and fewer couples, so it is unclear how much the difference in program success is related to household type or program quality. Lastly, the random synthetic data had exit probabilities unlike the other modes of data that are likely related to the random sampling nature of that data synthesis mode.

4.5 Limitations

Several limitations reduce the generalizability of results from this analysis. First, time, geographic, and historical biases in the administration of homeless services and the selection of families influences predictions of the programs associated with successful exits. These biases suggest that programs could operate very differently across jurisdictions. Secondly, this analysis only focused on last program and destination, and could miss the impact of utilizing services in a particular order that a trajectory-based approach would follow. Despite these limitations, the synthetic HMIS data used in this study are useful for comparing the clustering strategies that dominate homeless typology literature. The cluster trends (e.g., typically small groups or larger-sized clusters) across synthetic en-

vironments is an indicator of the robustness of a clustering approach. Lastly, the study did not address the uncertainty in random forest predictions. Recent research by Wager, Hastie, and Efron (2014) allows for the calculation of error bars with random forest classifiers and regressors. Future research will use Python implementation of random forest confidence intervals (Rokem et al. 2016)[9] to describe the uncertainty surrounding each prediction.

4.6 Conclusion

This study demonstrates an application of synthetic administrative data with with machine learning to understand questions in social science research. While the results show the kinds of answers that could be observed in actual data, the findings of this study do not reveal trends in historic data. With these caveats in mind, results from random forest classification trees revealed some variations in program success from pre- to post-Great Recession periods. The largest program difference across these periods was in rapid rehousing. The probability of other exit decreased from .52 in the pre-Great Recession period to .2 in the post-Great Recession era. Similarities across time periods include the high probability of other exit from emergency shelters and the relative successfulness of transitional and permanent supportive housing. While transitional housing had the highest probability of a successful program exit with subsidy in the pre-Great recession period, transitional and permanent supportive housing had near equal probabilities of successful exit in the post-Great Recession data.

Furthermore, several demographic differences in program success were observed in this study. Analysis revealed that program success was variable both across households and time periods. In both the pre-Great Recession and post-Great Recessions groups program type was more important for single-female households probabilities of successful exit than coupled households. For instance, couples had a near equal probability of other exit

[9] `https://github.com/scikit-learn-contrib/forest-confidence-interval`

in three post-Great Recession programs while single-female households were much more successful in transitional housing and to a lesser degree in permanent supportive housing. Within these household differences, racial differences also emerged where 100% of some racial minorities experienced other exits with emergency shelters but other racial minorities were able to achieve successful exits from shelters. The post-Great Recession period showed similar racial differences within households. These findings suggest that housing interventions should not only be tailored for households, but special attention to racial variation within households is necessary. Future research should assess the sensitivity of results to varying levels of noise and explore the basis of racial variations within similar household structures. This research highlights the granularity housing interventions may need to pursue by leveraging machine learning algorithms and synthetic modes of data to examine the robustness of housing program interventions in multiple eras.

4.6.1 References

Aiemagno, S. A. et al. 1996. "Assessing Substance Abuse Treatment Needs Among the Homeless: A Telephone-Based Interactive Voice Response System." *American Journal of Public Health* 86(11):1626–8. Retrieved May 11, 2017 (`http://ajph.aphapublications.org/doi/10.2105/AJPH.86.11.1626`).

Anderson, Isobel. 2003. "Synthesizing Homelessness Research: Trends, Lessons and Prospects." *Journal of Community & Applied Social Psychology* 13(2):197–205. Retrieved January 30, 2017 (`http://doi.wiley.com/10.1002/casp.721`).

Barrientos, Andrés F., Jerome P. Reiter, Ashwin Machanavajjhala, and Yan Chen. 2017. "Differentially Private Significance Tests for Regression Coefficients." *arXiv preprint arXiv:1705.09561*.

Bassuk, Ellen. 1984. "Is Homelessness a Mental Health Problem?" *American Journal of Psychiatry* 141(12):1546–50. Retrieved May 13, 2017 (`http://psychiatryonline.org/doi/abs/10.1176/ajp.141.12.1546`).

Bassuk, Ellen L. and Stephanie Geller. 2006. "The Role of Housing and Services in Ending Family Homelessness." *Housing Policy Debate* 17(4):781–806. Retrieved August 10, 2016 (`http://www.tandfonline.com/doi/abs/10.1080/10511482.2006.9521590`).

Bassuk, Ellen L. and Lynn Rosenberg. 1988. "Why Does Family Homelessness Occur? A Case-Control Study." *American Journal of Public Health* 78(7):783–88.

Bassuk, Ellen L. et al. 1997. "Homelessness in Female-Headed Families: Childhood and Adult Risk and Protective Factors." *American journal of public health* 87(2):241–48.

Bassuk, Ellen L., Carmela J. DeCandia, Corey Anne Beach, and Fred Berman. 2014. "America's Youngest Outcasts: A Report Card on Child Homelessness."

Bassuk, Ellen, Nicholas Huntington, Kim Lampereur, and Cheryl Amey. 2006. "Family Permanent Supportive Housing-Preliminary Research on Family Characteristics, Pro-

gram Models, and Outcomes - HUD Exchange." Retrieved May 14, 2017 (https://www.hudexchange.info/resource/951/family-permanent-supportive-housing-research-characteri

Bennett, Gary G., Melissa Scharoun-Lee, and Reginald Tucker-Seeley. 2009. "Will the Public's Health Fall Victim to the Home Foreclosure Epidemic?" *PLoS Medicine* 6(6):e1000087. Retrieved May 11, 2017 (http://dx.plos.org/10.1371/journal.pmed.1000087).

Breiman, Leo. 1996. "Bagging Predictors." *Machine learning* 24(2):123–40.

Breiman, Leo. 2001. "Random Forests." *Machine learning* 45(1):5–32.

Chawla, Shuchi, Cynthia Dwork, Frank McSherry, and Kunal Talwar. 2005. "On Privacy-Preserving Histograms." in *Uncertainty in Artificial Intelligence (UAI)*. Edinburgh, Scotland: Association for Uncertainty in Artificial Intelligence. Retrieved (https://www.microsoft.com/en-us/research/publication/on-privacy-preserving-histograms/).

Choi, Sam and Joseph P. Ryan. 2007. "Co-Occurring Problems for Substance Abusing Mothers in Child Welfare: Matching Services to Improve Family Reunification." *Children and Youth Services Review* 29(11):1395–1410. Retrieved August 10, 2016 (http://linkinghub.elsevier.com/retrieve/pii/S0190740907001211).

Cronley, C. and D. A. Patterson. 2012. "Does the Organization Matter? A Multilevel Analysis of Organizational Effects in Homeless Service Innovations." *Social Work Research* 36(1):70–79. Retrieved August 10, 2016 (http://swr.oxfordjournals.org/cgi/doi/10.1093/swr/svs020).

Culhane, Dennis P. and Stephen Metraux. 2008. "Rearranging the Deck Chairs or Reallocating the Lifeboats? Homelessness Assistance and Its Alternatives." *Journal of the American Planning Association* 74(1):111–21. Retrieved August 10, 2016 (http://www.tandfonline.com/doi/abs/10.1080/01944360701821618).

Culhane, Dennis P., Stephen Metraux, Jung Min Park, Maryanne Schretzman, and Jesse Valente. 2007. "Testing a Typology of Family Homelessness Based on Patterns of

Public Shelter Utilization in Four U.S. Jurisdictions: Implications for Policy and Program Planning." *Housing Policy Debate* 18(1):1–28. Retrieved January 30, 2017 (http://www.tandfonline.com/doi/abs/10.1080/10511482.2007.9521591).

Draine, Jeffrey, Mark S. Salzer, Dennis P. Culhane, and Trevor R. Hadley. 2002. "Role of Social Disadvantage in Crime, Joblessness, and Homelessness Among Persons with Serious Mental Illness." *Psychiatric Services* 53(5):565–73. Retrieved May 13, 2017 (http://psychiatryonline.org/doi/abs/10.1176/appi.ps.53.5.565).

Dwork, Cynthia. 2006. "Differential Privacy." Pp. 1–12 in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, vol. 4052. Venice, Italy: Springer Verlag. Retrieved (https://www.microsoft.com/en-us/research/publication/differential-privacy/).

Dwork, Cynthia. 2007. "Ask a Better Question, Get a Better Answer A New Approach to Private Data Analysis." Pp. 18–27 in *11th International Conference on Database Theory (ICDT 2007)*, vol. 4353. Barcelona, Spain: Springer. Retrieved (https://www.microsoft.com/en-us/research/publication/ask-a-better-question-get-a-better-answer-a-

Ellen, Ingrid Gould and Samuel Dastrup. 2012. "Housing and the Great Recession." *Policy Brief*.

Elsby, Michael, Bart Hobijn, and Aysegul Sahin. 2010. *The Labor Market in the Great Recession*. Cambridge, MA: National Bureau of Economic Research. Retrieved May 12, 2017 (http://www.nber.org/papers/w15979.pdf).

Fertig, Angela R. and David A. Reingold. 2008. "Homelessness Among at-Risk Families with Children in Twenty American Cities." *Social Service Review* 82(3):485–510. Retrieved May 13, 2017 (http://www.journals.uchicago.edu/doi/10.1086/592335).

Fisher, Benjamin W., Lindsay S. Mayberry, Marybeth Shinn, and Jill Khadduri. 2014. "Leaving Homelessness Behind: Housing Decisions Among Families Exiting Shelter."

Housing Policy Debate 24(2):364–86. Retrieved May 14, 2017 (http://www.tandfonline.com/doi/abs/10.1080/10511482.2013.852603).

Fowler, Patrick J. and Dina Chavira. 2014. "Family Unification Program: Housing Services for Homeless Child Welfare–Involved Families." *Housing Policy Debate* 24(4):802–14. Retrieved May 14, 2017 (http://www.tandfonline.com/doi/abs/10.1080/10511482.2014.881902).

Galea, Sandro and David Vlahov. 2002. "Social Determinants and the Health of Drug Users: Socioeconomic Status, Homelessness, and Incarceration." *Public health reports* 117(Suppl 1):S135.

Goodman, Sarena, Peter Messeri, and Brendan O'Flaherty. 2014. "How Effective Homelessness Prevention Impacts the Length of Shelter Spells." *Journal of Housing Economics* 23:55–62. Retrieved May 14, 2017 (http://linkinghub.elsevier.com/retrieve/pii/S1051137714000047).

Gould, Thomas E. and Arthur R. Williams. 2010. "Family Homelessness: An Investigation of Structural Effects." *Journal of Human Behavior in the Social Environment* 20(2):170–92. Retrieved April 8, 2017 (http://www.tandfonline.com/doi/abs/10.1080/10911350903269765).

Grant, Roy, Delaney Gracy, Grifin Goldsmith, Alan Shapiro, and Irwin E. Redlener. 2013. "Twenty-Five Years of Child and Family Homelessness: Where Are We Now?" *American Journal of Public Health* 103(S2):e1–e10. Retrieved April 8, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.2013.301618).

Hertzberg, Edwina L. 1992. "The Homeless in the United States: Conditions, Typology and Interventions." *International Social Work* 35(2):149–61.

Lee, Barrett A., Kimberly A. Tyler, and James D. Wright. 2010. "The New Homelessness Revisited." *Annual Review of Sociology* 36(1):501–21. Retrieved April 8, 2017 (http:

//www.annualreviews.org/doi/10.1146/annurev-soc-070308-115940).

Link, B. G. et al. 1994. "Lifetime and Five-Year Prevalence of Homelessness in the United States." *American Journal of Public Health* 84(12):1907–12.

Metraux, Stephen and Dennis P. Culhane. 1999. "Family Dynamics, Housing, and Recurring Homelessness Among Women in New York City Homeless Shelters." *Journal of family issues* 20(3):371–96.

Metraux, Stephen and Yi-Ping Tseng. 2017. "Using Administrative Data for Research on Homelessness: Applying a US Framework to Australia."

Metraux, Stephen, Magdi Stino, and Dennis P. Culhane. 2014. "Validation of Self-Reported Veteran Status Among Two Sheltered Homeless Populations." *Public Health Reports* 129(1):73–77.

Midderhoff, Latonia L. 1994. "The Student Perspective."

Miller, P. M. 2011. "A Critical Analysis of the Research on Student Homelessness." *Review of Educational Research* 81(3):308–37. Retrieved May 8, 2017 (http://rer.sagepub.com/cgi/doi/10.3102/0034654311415120).

Montgomery, Ann Elizabeth, Thomas H. Byrne, Daniel Treglia, and Dennis P. Culhane. 2016. "Characteristics and Likelihood of Ongoing Homelessness Among Unsheltered Veterans." *Journal of Health Care for the Poor and Underserved* 27(2):911–22. Retrieved May 14, 2017 (https://muse.jhu.edu/article/617504).

Nunez, Ralph and Cybelle Fox. 1999. "A Snapshot of Family Homelessness Across America." *Political Science Quarterly* 114(2):289–307. Retrieved April 8, 2017 (http://doi.wiley.com/10.2307/2657740).

Oberg, Charles N. 2011. "The Great Recession's Impact on Children." *Maternal and Child Health Journal* 15(5):553–54. Retrieved May 9, 2017 (http://link.springer.com/10.

1007/s10995-011-0807-8).

Page, Timothy and Roger M. Nooe. 2002. "Life Experiences and Vulnerabilities of Homeless Women: A Comparison of Women Unaccompanied Versus Accompanied by Minor Children, and Correlates with Children's Emotional Distress." *Journal of Social Distress and the Homeless* 11(3):215–31.

Pedregosa, Fabian et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(Oct):2825–30.

Pilkauskas, Natasha V., Janet M. Currie, and Irwin Garfinkel. 2012. "The Great Recession, Public Transfers, and Material Hardship." *Social Service Review* 86(3):401–27. Retrieved May 11, 2017 (http://www.journals.uchicago.edu/doi/10.1086/667993).

Ping, Haoyue, Julia Stoyanovich, and Bill Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." P. 42 in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Reiter, Jerome P. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18(4):531.

Rocha, Cynthia, Alice Johnson, Kay Young McChesney, and William Butterfield. 1996. "Predictors of Permanent Housing for Sheltered Homeless Families." *Families in Society: The Journal of Contemporary Social Services* 77(1):50–57.

Rossi, Peter H. 1991. *Down and Out in America: The Origins of Homelessness*. University of Chicago Press.

Rossum, Guido. 1995. *Python Reference Manual*. Amsterdam, The Netherlands, The Netherlands: CWI (Centre for Mathematics; Computer Science).

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of official Statistics*

9(2):461–68.

Sard, Barbara. 2009. "Number of Homeless Families Climbing Due to Recession." *Washington, DC: Center on Budget and Policy Priorities*.

Shinn, Marybeth. 1997. "Family Homelessness: State or Trait?" *American Journal of Community Psychology* 25(6):755–69. Retrieved April 8, 2017 (http://doi.wiley.com/10.1023/A:1022209028188).

Toro, Paul A. and David D. Wall. 1991. "Research on Homeless Persons: Diagnostic Comparisons and Practice Implications." *Professional Psychology: Research and Practice* 22(6):479–88. Retrieved May 14, 2017 (http://doi.apa.org/getdoi.cfm?doi=10.1037/0735-7028.22.6.479).

Treas, Judith. 2010. "The Great American Recession: Sociological Insights on Blame and Pain." *Sociological Perspectives* 53(1):3–18. Retrieved May 11, 2017 (http://spx.sagepub.com/lookup/doi/10.1525/sop.2010.53.1.3).

USICH, and VA, HUD. 2014. "National Alliance to End Homelessness: Core Components of Rapid Re-Housing." Retrieved May 19, 2017 (http://www.endhomelessness.org/library/entry/rapid-re-housing2).

Wager, Stefan, Trevor Hastie, and Bradley Efron. 2014. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife." *Journal of Machine Learning Research* 15(1):1625–51.

Walsh, Christine et al. 2014. "Permanent Supportive Housing for Families with Multiple Needs." Retrieved August 10, 2016 (http://www.homelesshub.ca/sites/default/files/Promising%20Practices%20for%20Homeless%20Families%20Final%20Report.pdf).

Washington, Donna et al. 2010. "Risk Factors for Homelessness Among Women Veterans." *Journal of Health Care for the Poor and Underserved* 21(1):82–91. Retrieved May 14, 2017 (http://muse.jhu.edu/content/crossref/journals/journal_of_health_care_for_the_

poor_and_underserved/v021/21.1.washington.html).

Wong, Yin-Ling Irene, Dennis P. Culhane, and Randall Kuhn. 1997. "Predictors of Exit and Reentry Among Family Shelter Users in New York City." *The Social Service Review* 441–62. Retrieved August 10, 2016 (http://www.jstor.org/stable/30012627).

Wood, D., R. B. Valdez, T. Hayashi, and A. Shen. 1990. "Homeless and Housed Families in Los Angeles: A Study Comparing Demographic, Economic, and Family Function Characteristics." *American Journal of Public Health* 80(9):1049–52. Retrieved April 8, 2017 (http://ajph.aphapublications.org/doi/10.2105/AJPH.80.9.1049).

Zhang, Wei and Oscar Gutierrez. 2007. "Information Technology Acceptance in the Social Services Sector Context: An Exploration." *Social Work* 52(3):221–31. Retrieved August 10, 2016 (http://sw.oxfordjournals.org/content/52/3/221.short).

Chapter 5

CONCLUSION

This dissertation shows that big data and computational social science can contribute to scholarly debates by creating respondent-driven ethnographic studies to examine demographic variations in opinions and leveraging machine learning to inform homelessness policy. Unstructured, user-generated digital trace data can be (inexpensively) gathered at scale and can form the basis of a digital ethnography to explore opinions. The ability to link multiple data sources, such as profile images to discern individual race/gender and location coordinates for socio-economic proxies, supports wide-ranging queries by contextualizing social media users. However, these digital data are not above reproach, and researchers must make judgments about the reliability and validity observations by establishing thresholds and filters for "usable data". While the findings from the Twitter study trend towards dispelling gender and racial differences in physical activity attitudes, interpretations should be cautious because of data scope and measurement concerns. For instance, this study did not use nationally representative data although communities under-represented in large scale surveys (e.g., minority men and women) are covered. Furthermore, both demographic measurement and opinion measurement are influenced by the inexact world of social media where users can choose any picture to represent themselves; sentiment analysis was not designed with the language of micro-blogging in mind.

Big data and computational social science may also be able to inform policy by examining an approach across multiple simulation environments. The second chapter bridges clustering approaches in homelessness typology research by examining two-dimensional

and family background clustering strategies across multiple types of synthetic data. Results suggest that homeless families' resource utilization is varied beyond the three-category typology favored by the two-dimensional clustering approach. The family background clustering approach was less restrictive in its categorization schema and favored classifying families in more groups to delineate resource utilization. As a result, it appears that homeless families have needs beyond the two-dimensional typology's characterization of episodic, transitional, or chronic homelessness. However, in some modes of synthetic data, the family background and atheoretical approaches which used 17 predictors and 19 predictors demonstrated the "curse of dimensionality" created by a larger predictor space. In these models, the clustering algorithm tended to detect 10 clusters to 20 clusters in stark contrast to the amount of clusters predicted from two predictors used in the two-dimensional approach. Together, the trend across modeling strategies and synthetic data types is that three groups do not adequately capture the diverse needs of homeless families. The challenge lies within navigating between the intractable approach of designing programs for 10 (or greater) types of homeless families and actual financial and human resources to address homelessness. This dissertation suggests that a human-in-the-loop approach that leverages machine and human strengths such as human aggregation and meta-synthesis of a cluster analysis (e.g., minority female cluster instead of a cluster for every female minority group) will be part of the solution to understanding homeless family needs.

The third chapter examines the relationship between homeless families' demographic backgrounds, housing program interventions, and probability of existing homelessness cycles pre- and post-Great Recession. This study also demonstrates how synthetic data can inform housing policy interventions by examining an approach in multiple simulations. The probability that a homeless family successful exited homeless cycles varied by household structure (e.g., single female, single male, couples) and sometimes household race within household structure. However, there are demographic differences in the composi-

tion of the pre- and post-Great Recession homeless families such as the distribution of household structures and sample size. While the housing program associated with best probability of a successful exit varied by synthetic data mode, the overarching takeaway is that different programs work better for some households (and racial groups) than others, and homeless service providers can investigate their real data for awareness of similar trends and prescribe policies.

In the future, big data and computational social science will strive to improve measurement accuracy by leveraging technology and traditional research. For instance, uncertainty about the demographic background of social media users can be removed by directly messaging users and asking them to take a census-style survey about relevant demographics. Additionally, social media sentiment analyses can be improved by assessing other micro-blogging features such as abbreviations, emojis, and the presence of intensifiers. Social media opinion mining research will benefit from the combination of a more refined sentiment analysis procedure with advancements in natural language processing to understand phrases such as puns.

Synthetic data will also play an important part in big data and computational social science research via result robustness and reproducibility. Result sensitivity to the noise component was not explored in these studies. Future research with synthetic data will compare methods across synthetic data types and varying levels of noise. Additionally, synthetic data can decrease the reproducibility crisis by allowing researchers to publish synthetic variants of their datasets (even sensitive data) and remove a layer of research opaqueness by allowing broader investigations of findings. Enhancements in data synthesizing methods and cryptography ensure that synthetic data will increase result robustness in and contribute to reproducible social science research. Hopefully, this dissertation provided a template for human-in-the-loop approaches to investigating health inequalities by applying emerging computational methods to multiple synthetic data sources (administrative data and social media) and discussing the limitations to the strategies considered.

Appendix A

CHAPTER 2 APPENDIX

A.0.1 Background

Examples of Physical Activity Survey Items

Affuso et al. (2011):

1) In order to relieve stress and maintain your health, how important is it for you personally to exercise—is it very important, somewhat important, not very important, or unimportant?; 2) In order to relieve stress and maintain your health, how important is it for you personally to get enough rest and relaxation—is it very important, somewhat important, not very important, or unimportant?; 3) Do you feel there are enough places in your neighborhood to be physically active, such as recreation centers, fitness centers, outdoor space, etc.?; 4) Do you think it is possible for a person to be overweight and still be healthy, or does being overweight mean a person is unhealthy?; 5) Do you agree or disagree with this statement: Exercise is necessary to be healthy.; 6) Do you think that being overweight can increase a person's risk of getting a disease like cancer, or not? Physical Activity participation—During the past month, other than your regular job, did you participate in any physical activities or exercise such as running, aerobics, golf, gardening, or walking for exercise? (pg. 3)

Bozionelos and Bennett (1999):

"Participants were required to indicate their level of agreement or disagreement

> with two statements (e.g., for me to participate in regular exercise during the next three weeks is . . . , etc.), using a seven-point Likert scale used on all the [Theory of Planned Behavior] items except intentions, on three bipolar adjective pairs for each statement (i.e. good/bad, harmful/beneficial and pleasant/unpleasant). The item responses were summed and divided by six to provide a total attitude score" (pg. 520)

Data Synthesizer Modes

In the random mode, ideal for the most sensitive data, the DataSynthesizer generates type consistent random variables, adding the highest degree of differential privacy. The correlated attribute mode uses a Bayesian network to calculate the relationship between variables and in cases where calculating a Bayesian network is too computationally expensive Ping, Stoyanovich, and Howe (2017) suggest using the independent attribute to sample from a noise-added distribution of the underlying data.

A.0.2 Example tweets

Example tweets:

Black women:

Positive: "Fast paced running ending with two hill repeats. Grueling but a good work out #Running #Training #WorkHard"

Negative: "Started my run wearing a jacket and I soon regretted it. #SummerStillHere #Running"

White women:

Positive: "My happy place! Thankful to be back doing what I love. #RUNNING #fitness #gym #love"

Negative: "I ran 20:29 with...#run #running fuck IBS I hate you"

Asian women:

Positive: "Enjoyed my time in nature today. #grateful #nature #family #walking"

Negative: "Since my knee injury I refuse to to anything that includes #running or walking long distance!"

Black men:

Positive: "Bodyweight #pullups will get you very strong #bodyweighttraining #equinox"

White men:

Positive: "Beautiful evening for a walk in the park #Love #Fitness #Walking"

Negative: "Running while sick sucks! Hope to sweat it out #running #keep-grinding #fitness"

Asian men:

Positive: "A beautiful day for a ride. #ironmantraining #biking #california #fitnesslifestyle"

Negative: "The struggle is real. #Running #Fitness #beFit #LifeStyleChange"

A.0.3 Activity-specific Subjectivity Scores

Table A.1: Activity-specific Subjectivity Scores

Race	Gender	Total Tweets	Hashtag	Hashtag Total	Subjectivity
Asian	Female	56	#walking	11	0.2
Asian	Male	157	#walking	1	0.01
Black	Female	100	#walking	3	0.03
Black	Male	155	#walking	11	0.07
Hispanic	Female	60	#walking	12	0.2
Hispanic	Male	200	#walking	33	0.16
Other	Male	5	#walking	0	0
White	Female	1889	#walking	243	0.13
White	Male	1905	#walking	176	0.09
Asian	Female	56	#running	44	0.79
Asian	Male	157	#running	101	0.64
Black	Female	100	#running	71	0.71
Black	Male	155	#running	129	0.83
Hispanic	Female	60	#running	28	0.47
Hispanic	Male	200	#running	150	0.75
Other	Male	5	#running	4	0.8
White	Female	1889	#running	1468	0.78
White	Male	1905	#running	1541	0.81
Asian	Female	56	#jogging	0	0
Asian	Male	157	#jogging	0	0
Black	Female	100	#jogging	0	0
Black	Male	155	#jogging	0	0

Race	Gender	Total Tweets	Hashtag	Hashtag Total	Subjectivity
Hispanic	Female	60	#jogging	4	0.07
Hispanic	Male	200	#jogging	8	0.04
Other	Male	5	#jogging	0	0
White	Female	1889	#jogging	23	0.01
White	Male	1905	#jogging	19	0.01
Asian	Female	56	#biking	0	0
Asian	Male	157	#biking	6	0.04
Black	Female	100	#biking	6	0.06
Black	Male	155	#biking	3	0.02
Hispanic	Female	60	#biking	4	0.07
Hispanic	Male	200	#biking	8	0.04
Other	Male	5	#biking	0	0
White	Female	1889	#biking	50	0.03
White	Male	1905	#biking	86	0.05
Asian	Female	56	#pushups	0	0
Asian	Male	157	#pushups	0	0
Black	Female	100	#pushups	0	0
Black	Male	155	#pushups	0	0
Hispanic	Female	60	#pushups	0	0
Hispanic	Male	200	#pushups	0	0
Other	Male	5	#pushups	0	0
White	Female	1889	#pushups	0	0
White	Male	1905	#pushups	0	0
Asian	Female	56	#pullups	0	0
Asian	Male	157	#pullups	0	0

Race	Gender	Total Tweets	Hashtag	Hashtag Total	Subjectivity
Black	Female	100	#pullups	0	0
Black	Male	155	#pullups	4	0.03
Hispanic	Female	60	#pullups	1	0.02
Hispanic	Male	200	#pullups	7	0.04
Other	Male	5	#pullups	1	0.2
White	Female	1889	#pullups	11	0.01
White	Male	1905	#pullups	40	0.02
Asian	Female	56	#homeworkouts	0	0
Asian	Male	157	#homeworkouts	0	0
Black	Female	100	#homeworkouts	0	0
Black	Male	155	#homeworkouts	0	0
Hispanic	Female	60	#homeworkouts	0	0
Hispanic	Male	200	#homeworkouts	0	0
Other	Male	5	#homeworkouts	0	0
White	Female	1889	#homeworkouts	0	0
White	Male	1905	#homeworkouts	0	0
Asian	Female	56	#bodyweightexercises	0	0
Asian	Male	157	#bodyweightexercises	0	0
Black	Female	100	#bodyweightexercises	0	0
Black	Male	155	#bodyweightexercises	0	0
Hispanic	Female	60	#bodyweightexercises	0	0
Hispanic	Male	200	#bodyweightexercises	0	0
Other	Male	5	#bodyweightexercises	0	0
White	Female	1889	#bodyweightexercises	0	0
White	Male	1905	#bodyweightexercises	1	0

Race	Gender	Total Tweets	Hashtag	Hashtag Total	Subjectivity
Asian	Female	56	#bodyweightworkouts	0	0
Asian	Male	157	#bodyweightworkouts	0	0
Black	Female	100	#bodyweightworkouts	0	0
Black	Male	155	#bodyweightworkouts	0	0
Hispanic	Female	60	#bodyweightworkouts	0	0
Hispanic	Male	200	#bodyweightworkouts	0	0
Other	Male	5	#bodyweightworkouts	0	0
White	Female	1889	#bodyweightworkouts	0	0
White	Male	1905	#bodyweightworkouts	0	0

Table A.2: Activity-specific Counts

Race	Gender	#walking	#biking	#running	#jogging
Asian	Female	11	0	44	0
Asian	Male	1	6	101	0
Black	Female	3	6	71	0
Black	Male	11	3	129	0
Hispanic	Female	12	4	28	4
Hispanic	Male	33	8	150	8
Other	Male	0	0	4	0
White	Female	243	50	1468	23
White	Male	176	86	1541	19

A.0.4 Activity-specific Polarity and Subjectivity Scores

Table A.3: #running-only Polarity Scores

Gender	Race	Polarity	Subjectivity	Hashtag
Female	Asian	0.02	0.79	#running
Female	Black	0.16	0.71	#running
Female	Hispanic	0.05	0.47	#running
Female	White	0.08	0.78	#running
Male	Asian	0.03	0.64	#running
Male	Black	0.01	0.83	#running
Male	Hispanic	0.06	0.75	#running
Male	Other	0	0.8	#running
Male	White	0.09	0.81	#running

Table A.4: #walking-only Polarity Scores

Gender	Race	Polarity	Subjectivity	Hashtag
Female	Asian	0.25	0.2	#walking
Female	Black	0.25	0.03	#walking
Female	Hispanic	0.15	0.2	#walking
Female	White	0.13	0.13	#walking
Male	Asian	0	0.01	#walking
Male	Black	0	0.07	#walking
Male	Hispanic	0.31	0.16	#walking
Male	Other	NA	0	#walking
Male	White	0.09	0.09	#walking

Table A.5: #jogging-only Polarity Scores

Gender	Race	Polarity	Subjectivity	Hashtag
Female	Asian	NA	0	#jogging
Female	Black	NA	0	#jogging
Female	Hispanic	0.25	0.07	#jogging
Female	White	0.04	0.01	#jogging
Male	Asian	NA	0	#jogging
Male	Black	NA	0	#jogging
Male	Hispanic	0.25	0.04	#jogging
Male	Other	NA	0	#jogging
Male	White	0.05	0.01	#jogging

Table A.6: #biking-only Polarity Scores

Gender	Race	Polarity	Subjectivity	Hashtag
Female	Asian	NA	0	#biking
Female	Black	0.17	0.06	#biking
Female	Hispanic	0	0.07	#biking
Female	White	0.11	0.03	#biking
Male	Asian	0	0.04	#biking
Male	Black	0.25	0.02	#biking
Male	Hispanic	0.25	0.04	#biking
Male	Other	NA	0	#biking
Male	White	0.08	0.05	#biking

Table A.7: #pullups-only Polarity Scores

Gender	Race	Polarity	Subjectivity	Hashtag
Female	Asian	NA	0	#pullups
Female	Black	NA	0	#pullups
Female	Hispanic	0	0.02	#pullups
Female	White	0.09	0.01	#pullups
Male	Asian	NA	0	#pullups
Male	Black	0	0.03	#pullups
Male	Hispanic	0.14	0.04	#pullups
Male	Other	0	0.2	#pullups
Male	White	0.03	0.02	#pullups

A.0.5 Sensitivity Analysis

Presentation of self

Table A.8: Mentions-only Polarity Scores

Gender	Race	Polarity
Female	White	0.07
Female	Black	0.17
Female	Hispanic	0.15
Female	Asian	0.14
Male	Asian	0
Male	Hispanic	0.14
Male	White	0.07
Male	Black	0.03

Table A.9: Retweets-only Polarity Scores

Gender	Race	Polarity
Female	White	0.09
Female	Black	0.17
Female	Asian	1
Male	White	0.16
Male	Hispanic	0
Male	Black	0

Table A.10: Original tweets-only Polarity Scores

Gender	Race	Polarity
Female	White	0.09
Female	Black	0.19
Female	Hispanic	0.03
Female	Asian	0.05
Male	Asian	0.04
Male	Black	0
Male	White	0.1
Male	Hispanic	0.1
Male	Other	0

Geographic filters

Table A.11: US Geolocated Tweets Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Black	0.17
Female	Hispanic	0.08
Male	Other	0
Female	Asian	0.07

Table A.12: Non-US Geolocated Tweets Polarity Scores

Gender	Race	Polarity
Female	White	0.11
Female	Asian	0.08
Female	Black	0.08
Male	White	0.09
Male	Black	0.12
Male	Asian	0.07

US geolocated tweets (SES proxy)

Table A.13: High income homes-only Polarity Scores

Gender	Race	Polarity
Female	White	0.19
Female	Asian	0
Female	Hispanic	0
Male	Asian	0
Male	White	0.13
Male	Hispanic	0

Table A.14: Middle income homes-only Polarity Scores

Gender	Race	Polarity
Female	White	0.08
Female	Black	0

Gender	Race	Polarity
Female	Hispanic	0
Male	White	0.05
Male	Hispanic	0
Male	Black	0
Male	Asian	0

Table A.15: Low income homes-only Polarity Scores

Gender	Race	Polarity
Female	White	0.18
Female	Asian	0
Male	White	0

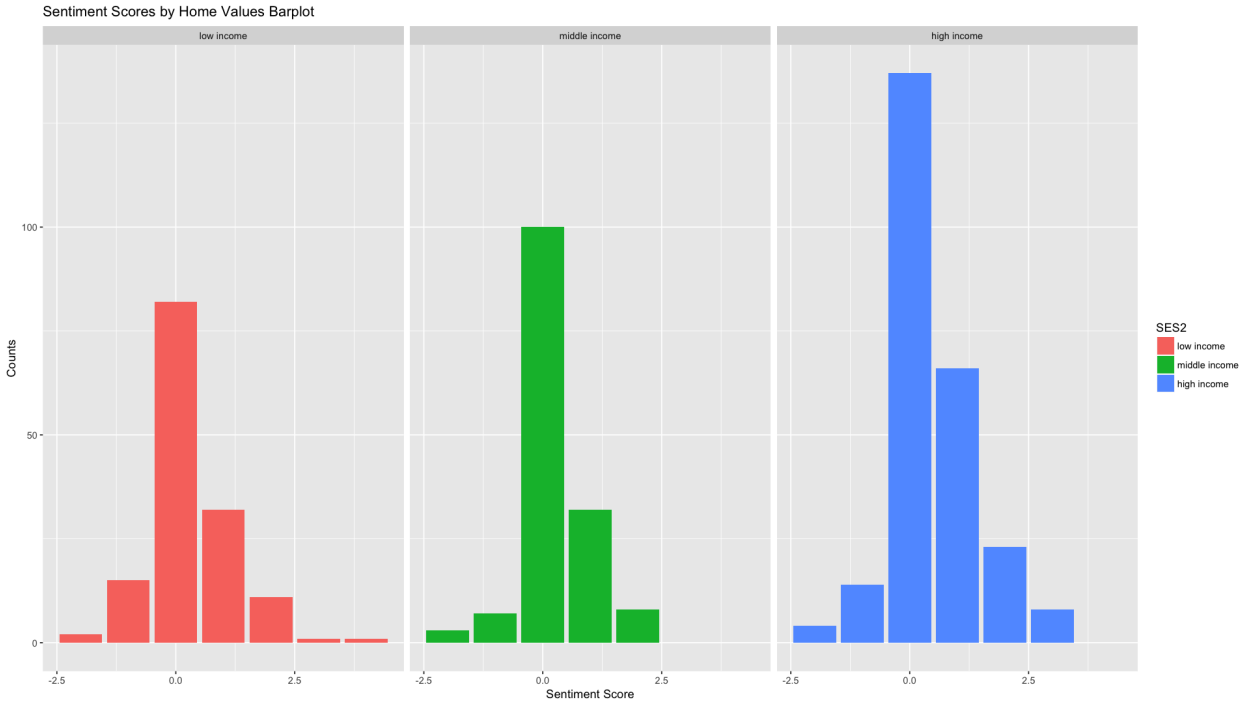


Figure A.1: Sentiment Score by House Values Histogram

User selectivity

Users with multiple tweets

Table A.16: Multiple tweets-only Polarity Scores

Gender	Race	Polarity
Female	White	0.08
Female	Black	0.16
Female	Hispanic	0.09
Female	Asian	0.06
Male	Asian	0.04
Male	Black	0.02

Gender	Race	Polarity
Male	White	0.09
Male	Hispanic	0.13
Male	Other	0

Users with one tweet

Table A.17: Single tweet users-only Polarity Scores

Gender	Race	Polarity
Female	White	0.12
Female	Black	0.21
Female	Hispanic	0.07
Female	Asian	0.1
Male	Asian	0
Male	Black	0.03
Male	Hispanic	0.08
Male	White	0.08
Male	Other	0

Subject reliability

Table A.18: Supplemental Hashtags: Intersectional Analysis of Polarity Scores

Gender	Race	Total	Proportion	Mean Age	Polarity
Male	White	50505	0.33	34.73	0.14

Gender	Race	Total	Proportion	Mean Age	Polarity
Female	White	72997	0.48	26.85	0.09
Male	Asian	5686	0.04	31.43	0.12
Female	Black	3131	0.02	33.02	0.11
Male	Black	12450	0.08	33.32	0.1
Female	Asian	7716	0.05	26.45	0.11

Demographic reliability

Racial reliability

Table A.19: 99th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Female	White	0.09
Male	White	0.1
Female	Black	0.16
Male	Hispanic	0.33
Female	Asian	0
Male	Black	0.01
Female	Hispanic	0.1
Male	Asian	0

Table A.20: 95th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Male	Asian	0.06
Female	White	0.09
Male	White	0.1
Male	Hispanic	0.19
Female	Black	0.19
Female	Asian	0.03
Male	Black	0.01
Female	Hispanic	0.08

Table A.21: 90th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Male	Asian	0.06
Female	White	0.08
Male	White	0.09
Male	Hispanic	0.15
Male	Black	0.02
Female	Black	0.18
Male	Other	0
Female	Asian	0.05
Female	Hispanic	0.14

Table A.22: 85th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Male	Asian	0.06
Female	White	0.08
Male	Black	0.01
Male	White	0.09
Male	Hispanic	0.12
Female	Black	0.16
Male	Other	0
Female	Asian	0.05
Female	Hispanic	0.12

Table A.23: 80th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Male	Asian	0.05
Female	White	0.08
Male	Black	0.01
Male	White	0.09
Male	Hispanic	0.12
Female	Black	0.17
Female	Hispanic	0.09
Male	Other	0
Female	Asian	0.05

Table A.24: 50th Percentile Racial Confidence Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Black	0.17
Female	Hispanic	0.08
Male	Other	0
Female	Asian	0.07

Gender reliability

Table A.25: 99th Gender Percentile Polarity Scores

Gender	Race	Polarity
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.13
Male	Asian	0.01
Male	White	0.09
Female	Hispanic	0.09
Male	Other	0
Female	Asian	0.07

Gender	Race	Polarity
Female	Black	0.14

Table A.26: 95th Gender Percentile Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Hispanic	0.09
Male	Other	0
Female	Asian	0.07
Female	Black	0.15

Table A.27: 90th Gender Percentile Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Hispanic	0.09
Male	Other	0

Gender	Race	Polarity
Female	Asian	0.08
Female	Black	0.14

Table A.28: 85th Gender Percentile Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Black	0.16
Female	Hispanic	0.09
Male	Other	0
Female	Asian	0.08

Table A.29: 80th Gender Percentile Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.08
Female	Black	0.16

Gender	Race	Polarity
Female	Hispanic	0.09
Male	Other	0
Female	Asian	0.08

Table A.30: 50th Gender Percentile Polarity Scores

Gender	Race	Polarity
Male	Asian	0.03
Female	White	0.09
Male	Black	0.02
Male	Hispanic	0.12
Male	White	0.09
Female	Black	0.17
Female	Hispanic	0.08
Male	Other	0
Female	Asian	0.07

Table A.31: Age Analysis of Polarity Scores

Gender	Race	Age Group	Polarity
Female	White	15_19	0.1
Female	Hispanic	15_19	0.17
Female	Black	15_19	0.75
Female	Asian	15_19	0
Female	White	20_24	0.09

Gender	Race	Age Group	Polarity
Female	Black	20_24	0
Female	Hispanic	20_24	0.08
Female	Asian	20_24	0
Female	White	25_29	0.09
Female	Black	25_29	0.16
Female	Asian	25_29	0.09
Female	Hispanic	25_29	0.08
Female	White	30_34	0.1
Female	Hispanic	30_34	0
Female	Asian	30_34	0.08
Female	Black	30_34	0.07
Female	White	35_39	0.09
Female	Hispanic	35_39	0
Female	Black	35_39	0.16
Female	White	40_44	0.02
Female	Hispanic	40_44	0
Female	White	45_49	0.07
Female	White	50_54	0.08
Female	White	55_59	0.09
Female	White	60_64	0.09
Female	White	65_69	0
Male	White	15_19	0.08
Male	Other	15_19	0
Male	Hispanic	15_19	0
Male	Black	15_19	0
Male	White	20_24	0.1

Gender	Race	Age Group	Polarity
Male	Hispanic	20_24	0.12
Male	Black	20_24	0
Male	Asian	20_24	0
Male	Hispanic	25_29	0.26
Male	White	25_29	0.1
Male	Other	25_29	0
Male	Asian	25_29	0
Male	Black	25_29	0
Male	Asian	30_34	0.05
Male	Black	30_34	0.03
Male	Hispanic	30_34	0.04
Male	White	30_34	0.12
Male	White	35_39	0.06
Male	Hispanic	35_39	0.08
Male	Black	35_39	0
Male	Asian	35_39	0.02
Male	White	40_44	0.07
Male	Hispanic	40_44	0.04
Male	Black	40_44	0
Male	White	45_49	0.04
Male	Black	45_49	0.01
Male	Asian	45_49	0
Male	Hispanic	45_49	0.27
Male	White	50_54	0.08
Male	Hispanic	50_54	0
Male	Asian	50_54	0

Gender	Race	Age Group	Polarity
Male	White	55_59	0.14
Male	Black	55_59	0.5
Male	Hispanic	55_59	0
Male	White	60_64	0.07
Male	Hispanic	60_64	0.29
Male	White	65_69	0.04

Table A.32: Pushups-only Sentiment Scores

Race	Gender	Mean	SD
Asian	Female	-0.07	1.16
Asian	Male	-0.16	1.18
Black	Female	0.39	1.03
Black	Male	0.05	1.11
White	Female	0.14	1.16
White	Male	-0.14	1.25

A.0.6 References

Affuso, Olivia, Tiffany L. Cox, Nefertiti H. Durant, and David B. Allison. 2011. "Attitudes and Beliefs Associated with Leisure-Time Physical Activity Among African American Adults." *Ethnicity & Disease* 21(1):63–67.

Bozionelos, G. and P. Bennett. 1999. "The Theory of Planned Behaviour as Predictor of Exercise: The Moderating Influence of Beliefs and Personality Variables." *Journal of Health Psychology* 4(4):517–29. Retrieved August 22, 2016 (`http://hpq.sagepub.com/`

cgi/doi/10.1177/135910539900400406).

Ping, Haoyue, Julia Stoyanovich, and Bill Howe. 2017. "DataSynthesizer: Privacy-Preserving Synthetic Datasets." P. 42 in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Appendix B

CHAPTER 3 APPENDIX

B.0.1 Appendix B: Goodness of Fit Tables

Table B.1: Two-dimensional Approach: Model BIC Comparison

	random 2D model BIC	independent 2D model BIC	correlated 2D model BIC
3 clusters	-25103	-4586	-6967
4 clusters	-24723	-1905	-6603
atheoretic clusters	99020	131810	170031

Table B.2: Family Background Approach: Model BIC Comparisons

	random background model BIC	independent background model BIC	correlated background model BIC
3 clusters	-109379	-53072	-40516
4 clusters	-108923	-34291	-35785
atheoretic clusters	-103266	-34291	-23530

B.0.2 Appendix B Summary Statistics: Synthesized Data by Mode

Table B.3: Summary Statistics: Synthesized Data by Mode

	Random Synthesized County	Independent Synthesized County	Correlated Synthesized County
single male	0.25	0.05	0.08
single female	0.24	0.77	0.74
couples	0.25	0.18	0.17
black	0.17	0.38	0.37
white	0.17	0.35	0.23
asian	0.16	0.13	0.04
multi-racial	0.17	0.07	0.08
native american	0.16	0	0.02
pacific islander	0.17	0.07	0.03
missing race	0.26	0	0
mean program duration (days)	3568	1257	451.9
mean programs count	4.04	1.53	1.46
mean substance abuse	1.53	0.34	0.07
mean physical disabilities	0.51	0.22	0.26
mean mental disabilities	0.51	0.09	0.23

	Random Synthesized County	Independent Synthesized County	Correlated Synthesized County
mean children	5.47	1.65	2.05
mean veteran status	0.5	0.02	0.08
mean parents	1.5	1.27	1.23
mean employed	0.51	0.4	0.47
mean benefits receipt	0.51	0.9	0.94
observations	4647	4647	2726

B.0.3 Appendix B: Two-dimensional Approach Tables

Table B.4: Two-dimensional Approach: Synthesized Random 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	3623	5568	1658	3528
mean programs (count)	1.26	4.03	3.96	6.76
mean employed	0.51	0.52	0.51	0.52
mean veteran status	0.5	0.5	0.5	0.51
mean substance abuse	1.57	1.54	1.52	1.49
mean children	5.39	5.57	5.44	5.46
mean benefits receipt	0.5	0.51	0.51	0.53
mean parents	1.5	1.5	1.49	1.51
mean physical disabilities	0.52	0.5	0.5	0.52
mean mental disabilities	0.52	0.5	0.5	0.52
single male	0.24	0.25	0.26	0.25
single female	0.24	0.24	0.24	0.26
couples	0.27	0.25	0.25	0.23
black	0.17	0.17	0.18	0.15
white	0.19	0.17	0.17	0.17
asian	0.15	0.16	0.17	0.16
multi-racial	0.16	0.16	0.17	0.17
native american	0.17	0.16	0.15	0.18
pacific islander	0.16	0.18	0.17	0.16
missing race	0.26	0.26	0.24	0.26
observations	834	1432	1502	897

Table B.5: Two-dimensional Approach: Synthesized Independent 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	3108	188.9	209	719.7
mean programs (count)	2.35	1	2	1
mean employed	0.39	0.4	0.38	0.42
mean veteran status	0.02	0.02	0.02	0.02
mean substance abuse	0.34	0.36	0.25	0.3
mean children	1.63	1.66	1.66	1.67
mean benefits receipt	0.91	0.89	0.89	0.91
mean parents	1.28	1.27	1.23	1.26
mean physical disabilities	0.23	0.21	0.23	0.22
mean mental disabilities	0.09	0.08	0.08	0.09
single male	0.04	0.05	0.06	0.04
single female	0.77	0.78	0.76	0.78
couples	0.19	0.17	0.18	0.18
black	0.37	0.37	0.4	0.39
white	0.35	0.36	0.32	0.35
asian	0.14	0.14	0.11	0.12
multi-racial	0.06	0.07	0.09	0.07
native american	0	0	0	0
pacific islander	0.09	0.06	0.08	0.07
missing race	0	0	0	0
observations	1563	1923	362	827

Table B.6: Two-dimensional Approach: Synthesized Correlated 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	144.8	1206	589.5	235
mean programs (count)	1.01	3.39	1.02	2.27
mean employed	0.47	0.46	0.48	0.43
mean veteran status	0.09	0.07	0.06	0.11
mean substance abuse	0.08	0.1	0.07	0.05
mean children	2.04	2.18	2	2.05
mean benefits receipt	0.94	0.91	0.96	0.94
mean parents	1.24	1.2	1.23	1.24
mean physical disabilities	0.26	0.26	0.26	0.26
mean mental disabilities	0.22	0.29	0.23	0.2
single male	0.08	0.09	0.07	0.07
single female	0.73	0.73	0.76	0.75
couples	0.18	0.17	0.16	0.16
black	0.38	0.29	0.38	0.44
white	0.23	0.3	0.23	0.18
asian	0.03	0.08	0.04	0.02
multi-racial	0.09	0.08	0.08	0.09
native american	0.02	0.03	0.02	0.03
pacific islander	0.02	0.05	0.03	0.04
missing race	0	0.01	0	0.02
observations	1189	391	909	245

B.0.4 Appendix B: Family Background Approach Tables

Table B.7: Family Background Approach: Synthesized Random 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	3577	3514	3569	3603
mean programs (count)	4.07	4.05	3.97	4.08
mean employed	0.49	0.55	0.52	0.5
mean veteran status	0.51	0.51	0.48	0.51
mean substance abuse	1.45	1.6	1.57	1.53
mean children	5.4	1.39	5.19	9.52
mean benefits receipt	0.51	0.56	0.51	0.48
mean parents	1.5	1.49	1.5	1.52
mean physical disabilities	0.51	0.59	0.48	0.47
mean mental disabilities	0.5	0.52	0.52	0.49
single male	0.25	0.21	0.25	0.27
single female	0.26	0.19	0.24	0.27
couples	0.23	0.3	0.26	0.23
black	0	0.19	0.31	0.18
white	0.35	0.16	0	0.19
asian	0	0.14	0.33	0.16
multi-racial	0.32	0.2	0	0.16
native american	0	0.1	0.35	0.17
pacific islander	0.33	0.21	0	0.14
missing race	0.25	0.3	0.25	0.23
observations	1502	758	1506	899

Table B.8: Family Background Approach: Synthesized Independent 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	1222	1283	1240	1270
mean programs (count)	1.6	1.55	1.51	1.51
mean employed	0.43	0.41	0.38	0.39
mean veteran status	0.02	0	0	0.05
mean substance abuse	2.99	0.05	0.04	0.1
mean children	1.73	1.65	1.64	1.63
mean benefits receipt	0.91	0.79	0.88	1
mean parents	1.28	1.28	1.26	1.26
mean physical disabilities	0.22	0.48	0.22	0.01
mean mental disabilities	0.08	0.07	0.05	0.13
single male	0.05	0.1	0.04	0.01
single female	0.78	0.5	0.77	0.98
couples	0.18	0.4	0.18	0.01
black	0.42	0.59	0	0.51
white	0.34	0	1	0.08
asian	0.13	0.19	0	0.19
multi-racial	0.06	0.11	0	0.1
native american	0	0	0	0
pacific islander	0.06	0.11	0	0.11
missing race	0	0	0	0
observations	436	1259	1351	1629

Table B.9: Family Background Approach: Synthesized Correlated 4-Cluster Demographics

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
mean duration (days)	418.5	507	479.2	438.3
mean programs (count)	1.43	1.52	1.5	1.45
mean employed	0.46	0.47	0.53	0.47
mean veteran status	0.07	0.08	0.05	0.08
mean substance abuse	0.01	0.01	2.49	0.01
mean children	2.09	2.06	2.04	2.02
mean benefits receipt	0.95	0.94	0.95	0.94
mean parents	1.25	1.25	1.18	1.21
mean physical disabilities	0.26	0.29	0.38	0.24
mean mental disabilities	0.22	0.21	0.32	0.24
single male	0.09	0.08	0.22	0.07
single female	0.75	0.75	0.59	0.75
couples	0.16	0.17	0.19	0.18
black	0	0	0.41	0.75
white	0	0.88	0.2	0
asian	0	0	0.07	0.08
multi-racial	0	0	0.05	0.17
native american	0.07	0	0.03	0
pacific islander	0	0.12	0.03	0
missing race	0	0	0	0
observations	642	704	74	1314

B.0.5 Appendix B: Demographic Background Variables

Danseco and Holden (1998)

> "Twelve variables were included in the cluster analysis representing housing problems and family characteristics: precipitating event for homelessness, history of homelessness which was assessed by a single item asking whether or not the family had ever been homeless prior to the current episode, number of moves, age of children, number of children, family type, parents' history of substance abuse, parents' welfare status, parents' physical and mental health status, PSI Total Stress, and PSI Life Stress score." (162)

B.0.6 References

Danseco, Evangeline R. and E. Wayne Holden. 1998. "Are There Different Types of Homeless Families? A Typology of Homeless Families Based on Cluster Analysis." *Family Relations* 47(2):159. Retrieved May 8, 2017 (`http://www.jstor.org/stable/585620?origin=crossref`).

Appendix C

CHAPTER 4 APPENDIX

C.0.1 Summary Statistics

Table C.1: pre-Great Recession Summary Statistics: Synthesized Data by Mode

	Random Synthesized Data	Independent Synthesized Data	Correlated Synthesized Data
single male	0.25	0.05	0.08
single female	0.25	0.77	0.67
couples	0.25	0.18	0.25
black	0.17	0.43	0.38
white	0.17	0.32	0.25
asian	0.16	0.13	0.04
multi-racial	0.17	0.06	0.08
native american	0.16	0	0
pacific islander	0.17	0.07	0.08
missing race	0	0	0
mean program duration (days)	3583	1230	639.5
mean programs count	4.07	1.47	2.17

	Random Synthesized Data	Independent Synthesized Data	Correlated Synthesized Data
mean substance abuse	1.52	0.27	0
mean physical disabilities	0.51	0.21	0.17
mean mental disabilities	0.51	0.1	0.21
mean children	5.56	1.68	2.79
mean veteran status	0.51	0.02	0.08
mean parents	1.5	1.27	1.17
mean employed	0.51	0.41	0.33
mean benefits receipt	0.51	0.91	0.92
observations	3146	629	24

Table C.2: post-Great Recession Summary Statistics: Synthesized Data by Mode

	Random Synthesized Data	Independent Synthesized Data	Correlated Synthesized Data
single male	0.25	0.05	0.08
single female	0.23	0.77	0.74
couples	0.26	0.18	0.17
black	0.17	0.37	0.37
white	0.18	0.36	0.23
asian	0.16	0.13	0.04
multi-racial	0.16	0.07	0.08
native american	0.17	0	0.02
pacific islander	0.16	0.07	0.03
missing race	0	0	0
mean program duration (days)	3538	1261	450.3
mean programs count	3.97	1.54	1.46
mean substance abuse	1.54	0.35	0.08
mean physical disabilities	0.5	0.22	0.26
mean mental disabilities	0.5	0.08	0.23
mean children	5.28	1.65	2.04
mean veteran status	0.5	0.02	0.08

	Random Synthesized Data	Independent Synthesized Data	Correlated Synthesized Data
mean parents	1.5	1.27	1.23
mean employed	0.52	0.39	0.47
mean benefits receipt	0.52	0.9	0.94
observations	1501	4018	2702

Table C.3: Entire Data Summary Statistics: Synthesized Data by Mode

	Random Synthesized County	Independent Synthesized County	Correlated Synthesized County
single male	0.25	0.05	0.08
single female	0.24	0.77	0.74
couples	0.25	0.18	0.17
black	0.17	0.38	0.37
white	0.17	0.35	0.23
asian	0.16	0.13	0.04
multi-racial	0.17	0.07	0.08
native american	0.16	0	0.02
pacific islander	0.17	0.07	0.03
missing race	0.26	0	0
mean program duration (days)	3568	1257	451.9
mean programs count	4.04	1.53	1.46
mean substance abuse	1.53	0.34	0.07
mean physical disabilities	0.51	0.22	0.26
mean mental disabilities	0.51	0.09	0.23
mean children	5.47	1.65	2.05

	Random Synthesized County	Independent Synthesized County	Correlated Synthesized County
mean veteran status	0.5	0.02	0.08
mean parents	1.5	1.27	1.23
mean employed	0.51	0.4	0.47
mean benefits receipt	0.51	0.9	0.94
observations	4647	4647	2726

C.0.2 Program Success Tables

Program Success: Synthetic Correlated Data

pre-Great Recession

Table C.4: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.38	0.18	0.44
Transitional Housing	0.4	0.1	0.5
Emergency Shelter	0.35	0.55	0.1
Permanent Housing	0.81	0.04	0.15

post-Great Recession

Table C.5: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.45	0.32	0.24
Transitional Housing	0.36	0.32	0.32
Emergency Shelter	0.45	0.38	0.18
Permanent Housing	0.38	0.41	0.21

Entire Data

Table C.6: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.36	0.21
Transitional Housing	0.36	0.32	0.32
Emergency Shelter	0.45	0.35	0.2
Permanent Housing	0.38	0.41	0.21

Program Success: Synthetic Independent Data

pre-Great Recession

Table C.7: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.52	0.41	0.07
Transitional Housing	0.31	0.49	0.2
Emergency Shelter	0.25	0.62	0.13
Permanent Housing	0.45	0.26	0.3

post-Great Recession

Table C.8: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.4	0.19
Transitional Housing	0.4	0.39	0.22
Emergency Shelter	0.69	0.19	0.11
Permanent Housing	0.5	0.32	0.18

Entire Data

Table C.9: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.42	0.42	0.16

	Other	Successful	Successful with Subsidy
Transitional Housing	0.39	0.39	0.21
Emergency Shelter	0.67	0.21	0.12
Permanent Housing	0.51	0.31	0.19

Program Success: Synthetic Random Data

pre-Great Recession

Table C.10: pre-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.74	0.1	0.16
Transitional Housing	0.67	0.16	0.17
Emergency Shelter	0.7	0.14	0.17
Permanent Housing	0.7	0.15	0.15

post-Great Recession

Table C.11: post-Great Recession Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.65	0.24	0.11
Transitional Housing	0.6	0.12	0.28
Emergency Shelter	0.57	0.19	0.23
Permanent Housing	0.65	0.17	0.17

Entire Data

Table C.12: Entire Data: Exit Probabilities

	Other	Successful	Successful with Subsidy
Rapid Rehousing	0.71	0.14	0.15

	Other	Successful	Successful with Subsidy
Transitional Housing	0.67	0.14	0.19
Emergency Shelter	0.64	0.15	0.2
Permanent Housing	0.67	0.16	0.17

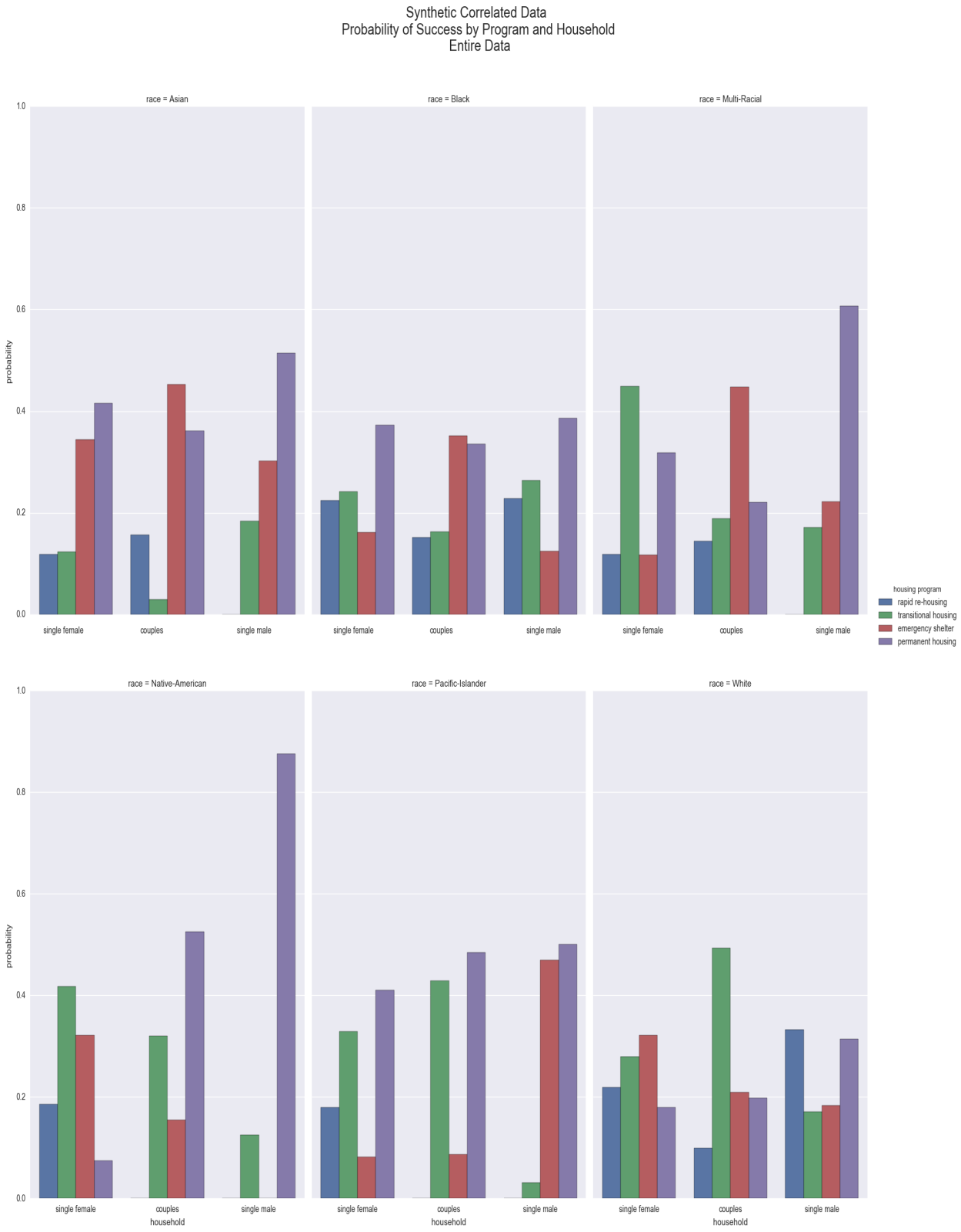


Figure C.1: Correlated Entire Data

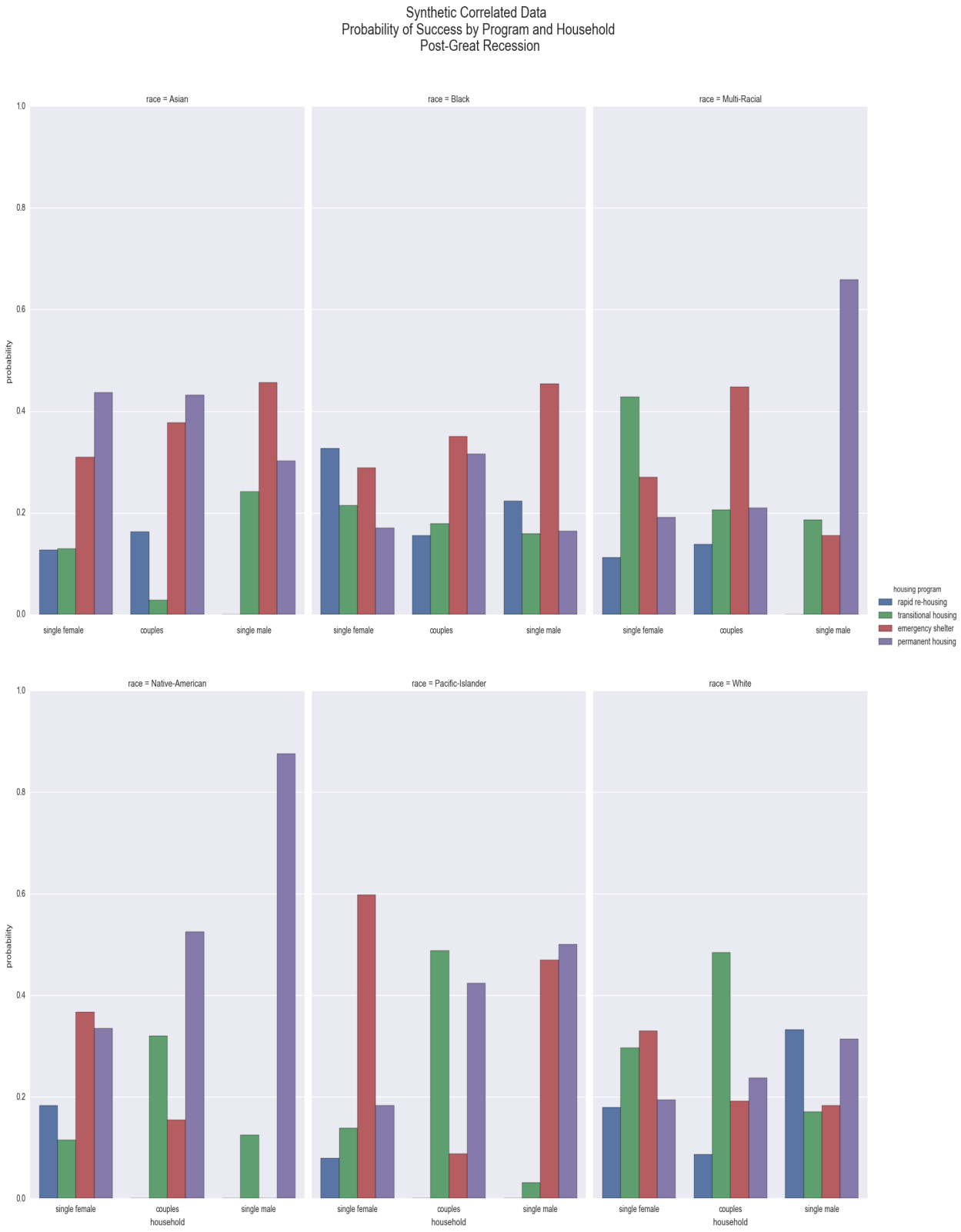


Figure C.2: Correlated post-Recession Data

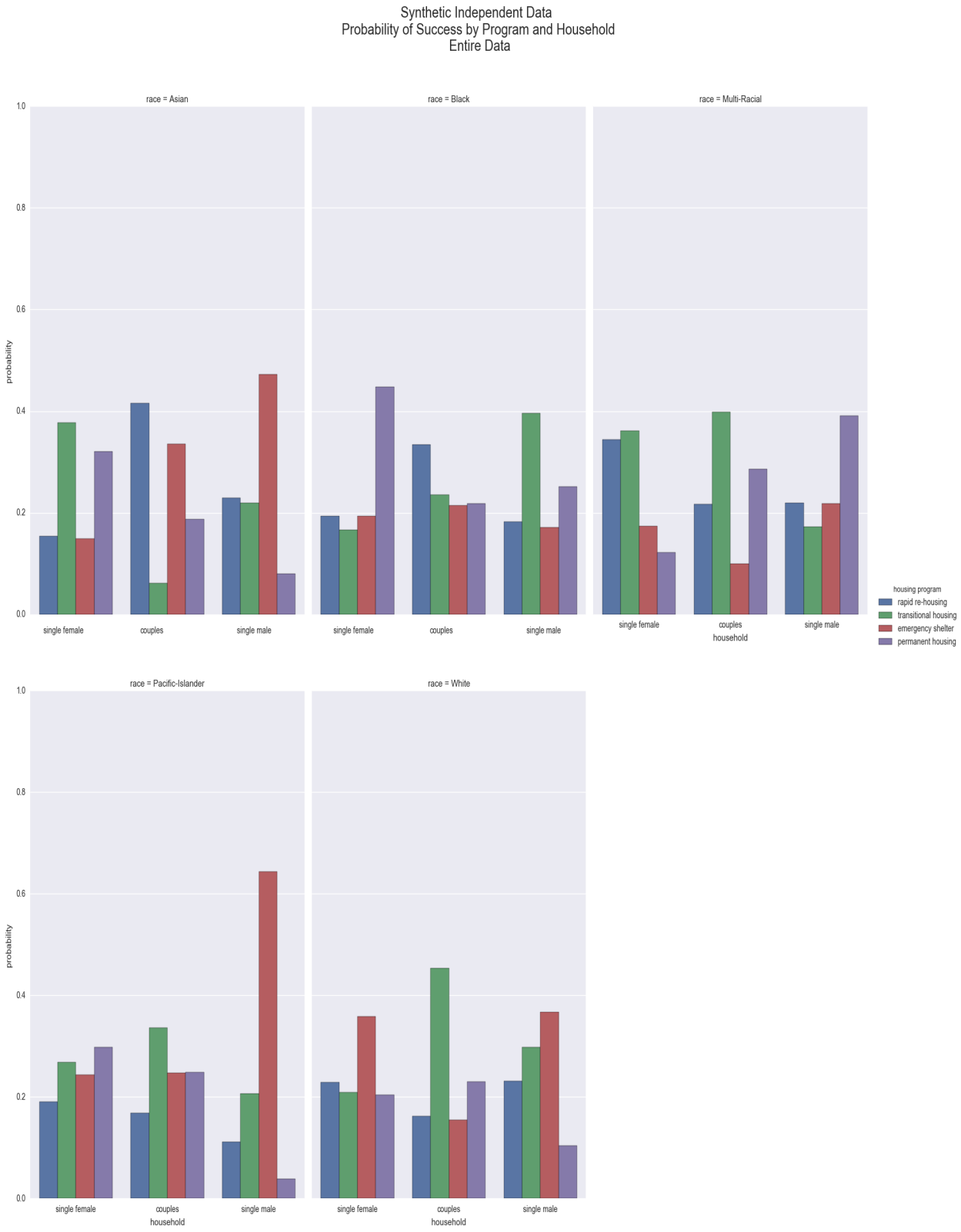


Figure C.3: Independent Entire Data

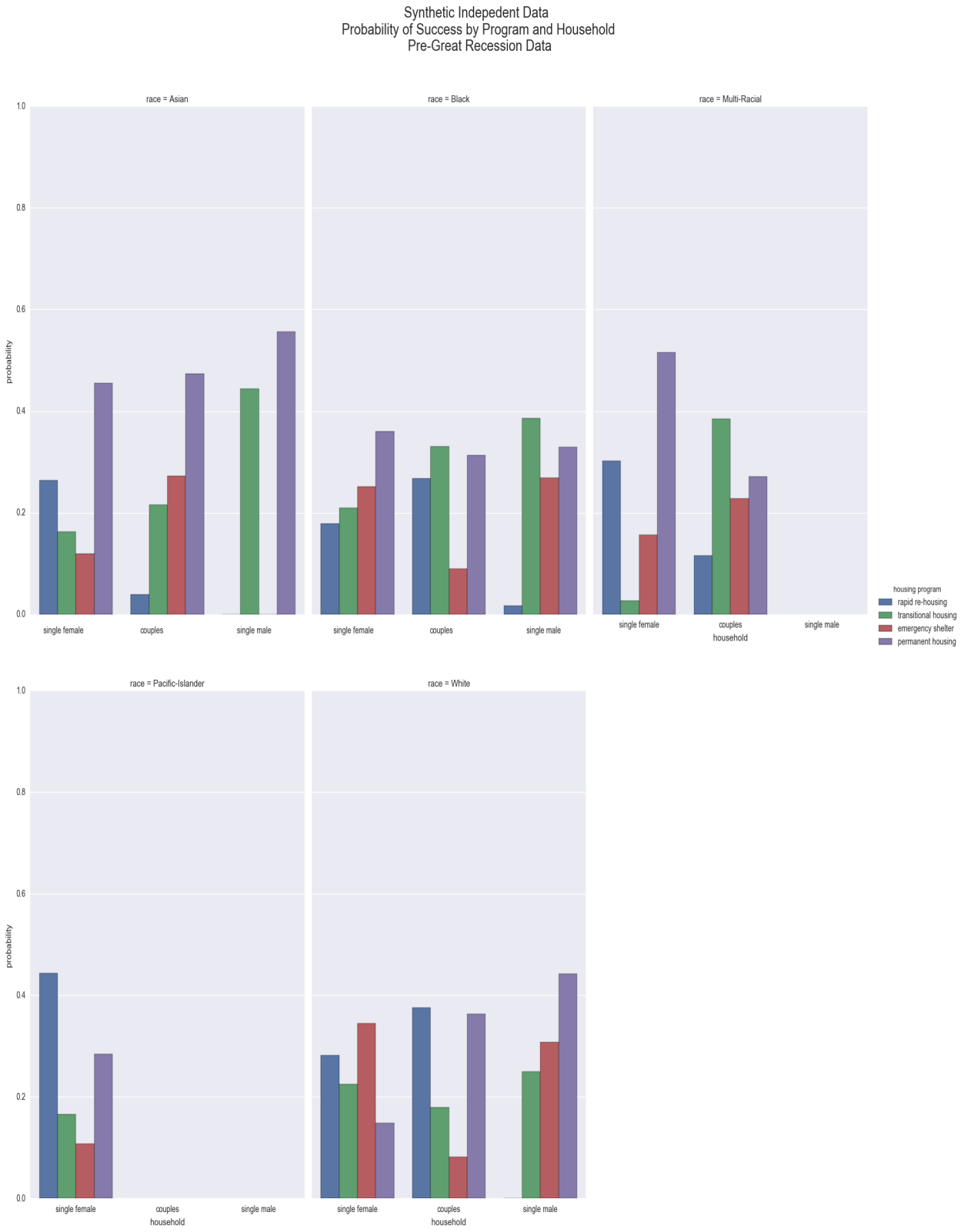


Figure C.4: Independent pre-Great Recession Data

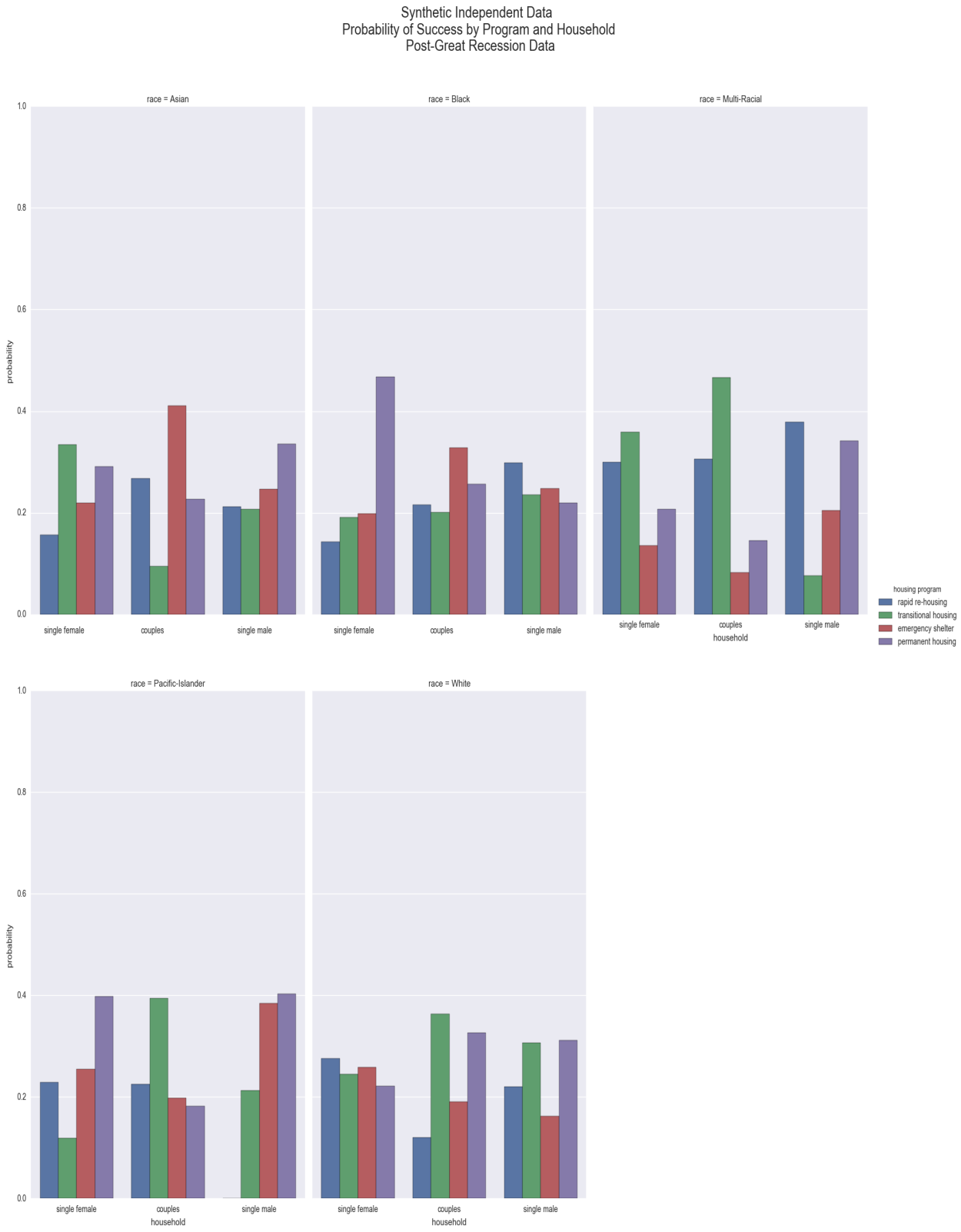


Figure C.5: Independent post-Great Recession Data

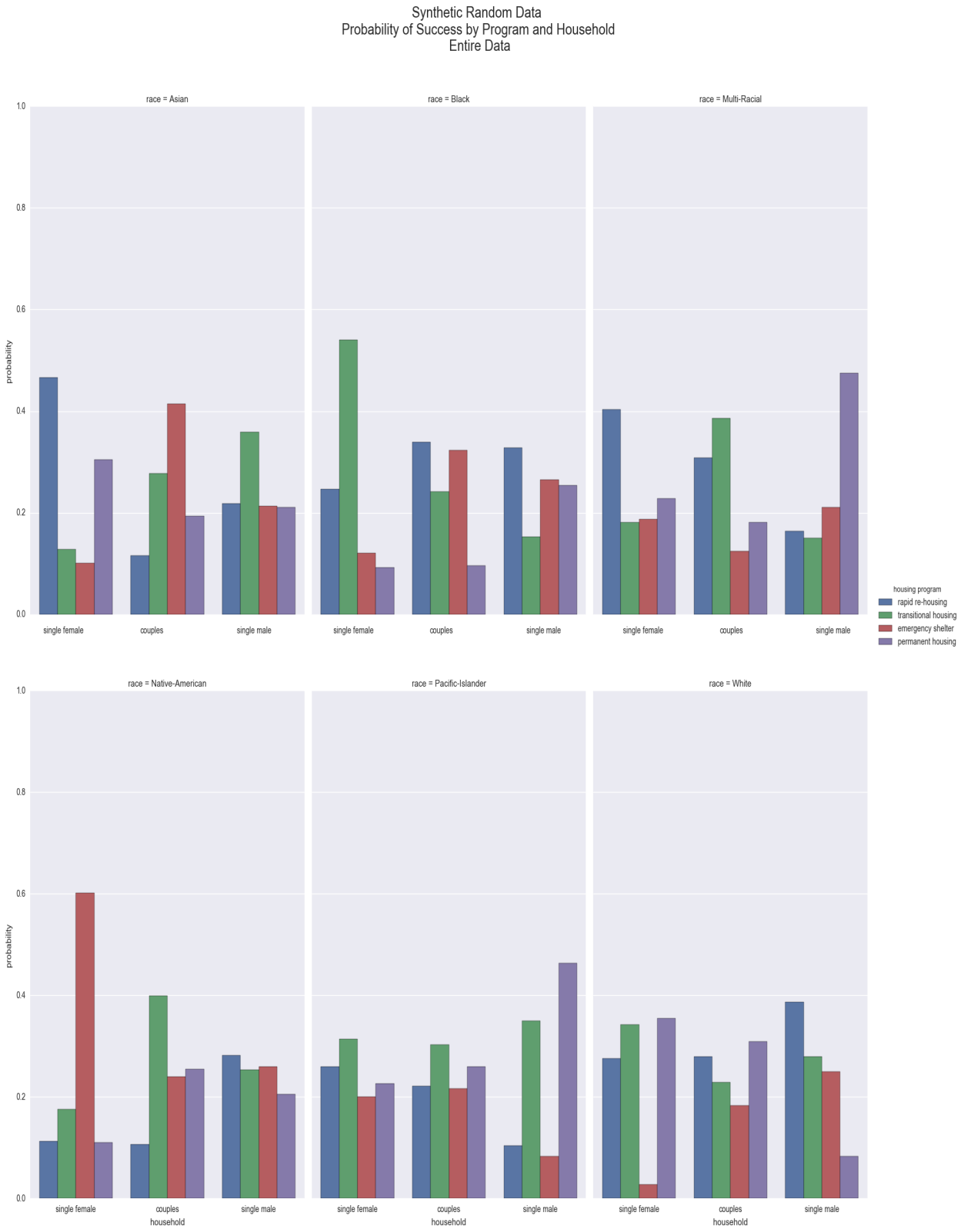


Figure C.6: Random Entire Data

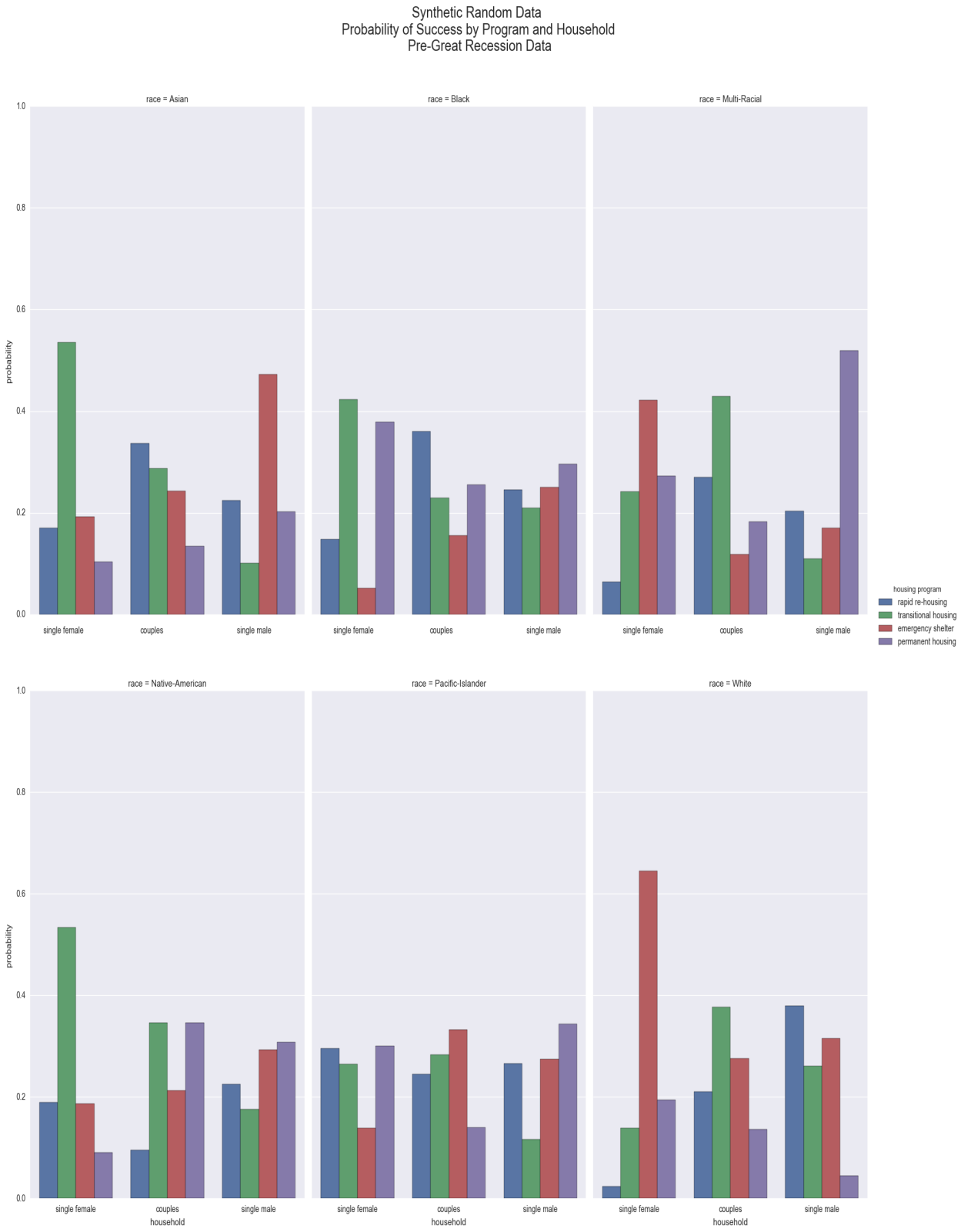


Figure C.7: Random pre-Great Recession Data

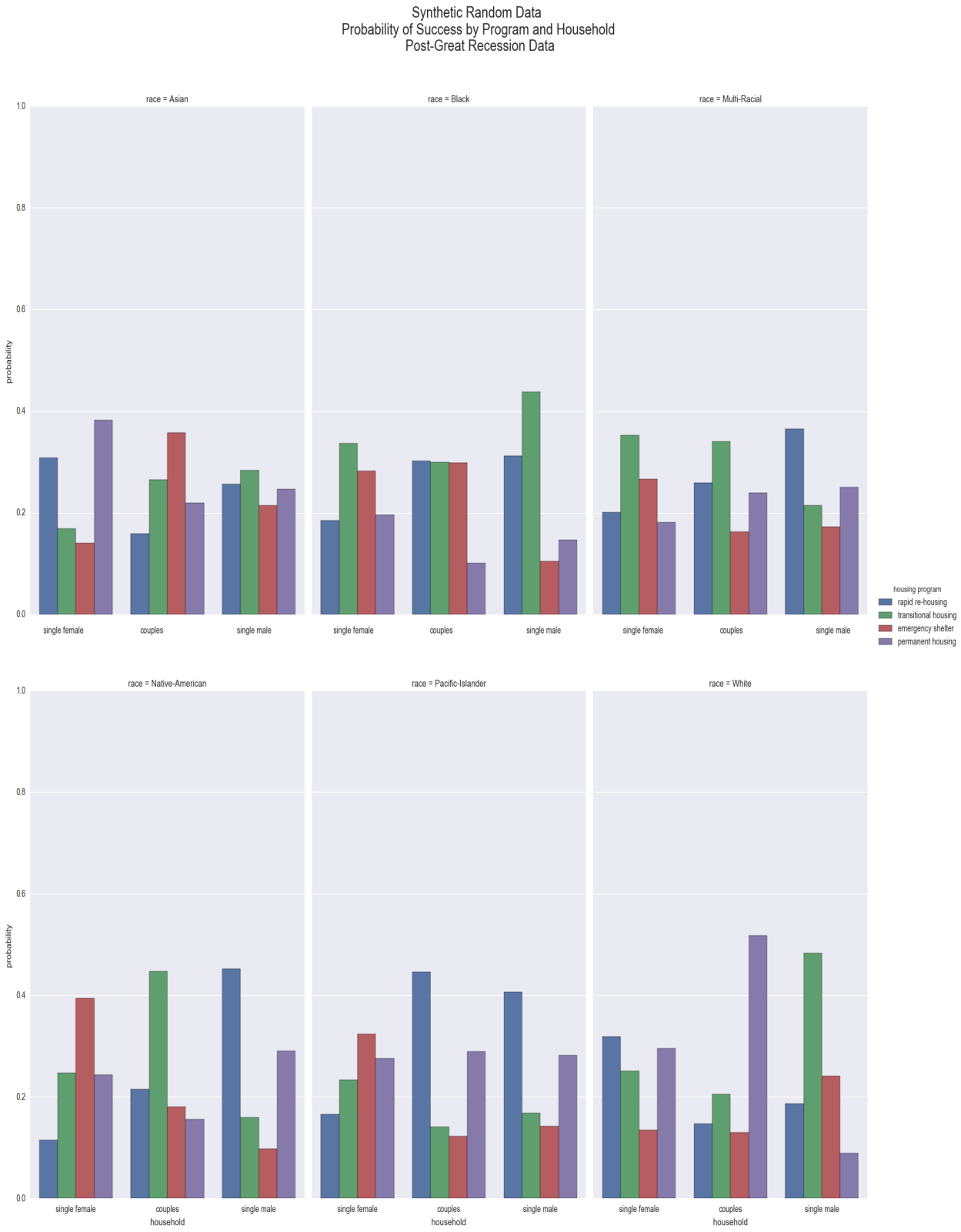


Figure C.8: Random post-Great Recession Data

VITA

Kivan Polimis is a computational social scientist focused on health development and applying statistical techniques to investigate disparities in health care, transportation, and the legal system. Recently, Kivan has benefited tremendously from training by multiple computational social science communities including the UW's Data Science for Social Good, the Russell Sage Foundation's Summer Institute in Computational Social Science, and the Department of Veteran Affairs' Big Data-Scientist Training Enhancement Program. In October 2017, he will begin a postdoctoral fellowship at Università Bocconi's Dondena Centre for Research on Social Dynamics and Public Policy in Milan.