NORTHWESTERN UNIVERSITY

Identifying and Addressing Structural Inequalities in the Representativeness of Geographic Technologies

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Isaac L. Johnson

EVANSTON, ILLINOIS

March 2019

ABSTRACT

Historically, there have been large disparities in the degree to which different communities have access to resources and representation within society. With the increased availability of the internet and the growth of user-generated content platforms like Twitter and Wikipedia, there are opportunities to alleviate some these long-standing barriers to access and representation. However, there is growing evidence that many of these technologies may instead be reinforcing some of these long-standing disparities. In the first part of this dissertation, we examine how different segments of the population are represented in social media and peer production, with a particular focus on the urban-rural divide. We demonstrate that it is important to go beyond surveys of participation rates, that online representation must be evaluated in the context of different consumers of online content. Across three studies, we find that even with proportionate participation in rural areas, disparities in online representation can still remain in the quality of content viewed by users, robustness of conclusions in computational social science about these areas, and precision of algorithms that are trained from this online data. In the second part of this dissertation, we focus on the domain of geographic algorithms, evaluating how biases arise in vehicle routing, place recommendation, and geographic representation learning. We develop a framework for choosing geographic hyperparameters that affect the performance of these technologies. We provide methods

for evaluating the fairness of these technologies with regards to these geographic hyperparameters. Across these studies, we find a complicated relationship between choices made in designing the algorithms underlying these technologies and the impact of these algorithms on communities. I conclude with an overview of best practices for working with geographic data and human-centered algorithms with the goal of developing technologies that are more equitable and more readily evaluated for disparities in their impact on different communities.

Acknowledgments

It goes without saying that the amazing opportunities of the last several years would not have occurred without the unfailing support of many. I offer a few words here but know that they are incomplete. I'm sure I left out a few of you but not by design.

First of all, I would like to thank my advisor Brent Hecht. You took a chance with me, as someone who largely lacks the traditional qualifications of someone seeking a computer science PhD, and were patient, supportive, and respectful in our many many discussions and collaborations. I have learned much over these years as a direct result of your mentorship and have been incredibly grateful that throughout all of it, I have been able to openly disagree and talk through all manner of challenges while always agreeing about the truly important matters.

Thank you to the rest of my committee. Nell O'Rourke for the incredibly insightful questions and new perspectives you brought to this work. Shilad Sen for your advice on these final projects and pushing me to really define what I was saying. Darren Gergle for managing to do just about everything and still making time for coffee, chats, and overseeing a dissertation.

Thanks to my family—I couldn't ask for a better one. Thanks to Jeanine for being beyond supportive through the end. Thanks to my Pittsburgh people for encouraging me to give this a shot.

As well, a long line of unbelievable roommates (Jim and Sarah especially for welcoming me to Chicago, Rory, Diana, Mitch, Stevie, Andrea, Hayk, Scott, Patrick), loving pets (Merle, Smokey, Dragon, Tosci, Ari), and ever-supportive friends (Max, Max, Max, Joe, Rebecca, Alex, Elias, Jerry, Gabe, Ephraim, Joanna, Lina, Laura, Sarah, Jake, Jes, Scott, Mark, Mollie, Janine, and many more).

Thank you to the University of Minnesota and GroupLens for welcoming me into the world of computer science and then continuing to be supportive and flexible with the shift to Northwestern. Thank you to PSA Research, to Allen, Hanlin, and Nick for putting up with my idiosyncrasies without complaint. To CollabLab, the many folks across HCI, the Knight Lab, and Northwestern University for accepting me into the fold and being a great second home in which to finish up my dissertation. To Marcia for the many chats and Maria for never getting frustrated and helping me navigate my ever-changing status.

Thank you to Google for two incredible internships, my mentors (Kacey and Chris), my teams (Ariel, Dave, Alyssa, Aidan, and many others). Kacey, I learned so much and smiled so much doing it. And Chris, for the jazz, encouraging me to walk those nine flights of stairs, and continued mentorship and perspective.

And finally, thanks to my new Wikimedia family for granting me this next opportunity! You make all of this so worth it.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The research contained within this dissertation proposal is motivated by concerns around inequality, internet technologies, and how these two intersect. Despite the promises of the World Wide Web to reduce existing societal inequalities [**153, 304, 89, 88, 163, 205**], there are concerns that internet technologies and big data have instead largely mirrored and, in certain cases, amplified these disparities [**19, 72, 234**]. Just as the US government practice of redlining (race-based denial of services to neighborhoods) contributed to racial inequality [**2**], there is evidence that analogous discrimination is occurring via "technological redlining" [**232**] (data and algorithms that reinforce historical disparities). With this history in mind, my goals are to 1) better understand how different communities of people are represented online, especially in the context of research and as data for algorithmic systems, and, 2) design more responsible algorithms to account for biases in these representations.

In this thesis, the concept of online representation is predicated on the belief that if a given community puts some amount of effort into an online platform or technology, then that community should receive the same level of benefits from that technology as any other community that put in that level of effort. To understand the importance and complexities of online representation, consider a population such as rural Americans. To what degree can they access high-quality content about their towns on Wikipedia or OpenStreetMap? Can computational social science researchers accurately study these communities through social media trace data? Do place recommendation technologies or language models perform

as well in these areas as they do in urban regions? And then, for populations that lack robust representation online, how can we build algorithmic technologies that correct for this disadvantage?

Inequality, be it in terms of opportunity, income, or political representation, is a fundamental problem faced by society [**89**]. There is hope, though, that the World Wide Web, big data, and associated technologies would help to address some of these disparities in access to resources and opportunity [**153**]. While humans struggle with discrimination and biases, computational models might be trained to be more "objective". With internet access, anyone could contribute to distributed knowledge sources like Wikipedia or social discourse as with Twitter.

Instead, evidence has been building that participation online varies heavily across demographic lines—e.g., population density, race, income, education. And while some of these gaps have been closing in countries like the United States [**39**], there are still huge gaps in countries without the same resource advantages [**17**]. We find across the first half of this dissertation, however, that achieving equitable online technologies is more complicated than achieving equal participation or coverage. Just as spatial homophily or gerrymandering can reduce the value of a populations electoral vote regardless of turnout, it is important to understand the context around how online data is used when quantifying representation and designing technologies to mitigate biases in this representation.

Specifically, across Part I of this thesis, I explore how well different online platforms represent urban and rural populations in the context of users, research, and algorithms. In Chapters 3 and 4, I discuss two studies that examine multiple social media and peer production platforms. I demonstrate that bots and non-local contributors often produce a disproportionate amount of content in rural areas, leading to low-quality content for users

and misleading research conclusions. Within the context of algorithms, I show in Chapter 5 that correcting for lower online participation rates in rural areas is not always sufficient for achieving parity in algorithmic performance. Structural factors such as low population density and differences in online behavior, which cannot be remedied by rebalancing data alone, pose major obstacles to achieving equitable algorithmic performance for these communities.

The emergence of algorithmic technologies, with their large potential for (disparate) impact [**234, 19**], brings special urgency to the question of geographic biases in representation. Part I demonstrates that we cannot simply hope that greater participation will correct for disparities in the benefits of the technologies. Geographic technologies such as place recommendation [**198**], vehicle routing [**344**], and even games such as Pokémon Go [**48**] are widespread and guide important economic activity (see §2.4). Core to ensuring that these technologies do not simply reproduce or exacerbate existing inequalities is the ability to audit the impact of these technologies and make explicit their biases so that they might be addressed.

In Part II of this thesis then, I focus on how we can design more responsible geographic technologies to account for biases in representation. I focus on evaluating the fairness of two specific geographic technologies, vehicle routing (§7) and place recommendation (§8), as well as the more general challenge of understanding what is encoded within large-scale geographic trace data upon which many algorithms are trained (§9). For each of these studies, I develop methods for evaluating how different choices in the design of the algorithm affect what it encodes and the distribution of its impact. I develop metrics that show that alternative vehicle routing algorithms can substantially redistribute traffic, but that the the neighborhoods that receive increased or decreased traffic can be difficult to predict a priori (§7). I demonstrate that embeddings learned by collaborative-filtering algorithms for place

recommendation strongly encode location but that removing that bias does not lead directly to more equitable recommendations (§8). Finally, I provide general methods for representing and evaluating what is encoded by large-scale, unstructured geographic data (§9).

Together, these findings provide valuable insight into the representativeness of geographic user-generated content, most specifically in how it might be used within the context of algorithms. This includes best practices for how to more effectively incorporate this content into geographic technologies as well as guidance for how we might evaluate these technologies to ensure that their benefits are distributed more evenly. The hope is that the frameworks and results that I provide might guide algorithm designers and the broader public in discussions about the impact of these technologies, what they should optimize, and how we might design them to not merely encode structural inequalities within society.

CHAPTER 2

Related Work

2.1. A Note on Terminology

User-Generated Content (UGC) is a common term that applies to a wide variety of content—e.g., posts to social media sites like tweets, photos, or place check-ins, edits to peer-production platforms like Wikipedia or OpenStreetMap, reviews on platforms like Yelp or TripAdvisor. This is also called human-generated content; this is an attempt to emphasize that this data comes from actual people, as opposed to the abstracted "user". I will retain UGC, largely because it is still the most common usage. The explicitly spatial subset of UGC is often called Volunteered Geographic Information (VGI). While this term ignores that much of this geographic UGC is by no means "volunteered" (e.g., trace data that many do not understand is being collected) [**132**], it conveniently encapsulates both the geographic component and UGC component. As such, I use both UGC and VGI in several chapters as concise descriptors.

Regarding user-generated content, "participation" and "coverage" are different measures of the quantity and distribution of content. "Participation" refers to the proportion of a given population that contributes to an online platform. This can be determined via surveys or from empirical studies—e.g., counting the unique number of users who post content in a given area and normalizing that by the population of that area. "Coverage" refers to the amount of content about a given topic or area. It can be normalized by the population (e.g., tweets per capita), area (e.g., tweets per square kilometer), or some ground-truth metric

(e.g., proportion of roads with speed limit tags on OpenStreetMap). While it relates to participation, there is not always a direct connection as it is affected users' activity levels, the degree to which there are outside contributors, and, when there is ground truth, the amount of content to be described.

The phrases "algorithm", "computational model", and "intelligent technology" are used somewhat interchangeably throughout this work. Though the terms do not all have the same meaning, there is a fair bit of overlap. Most importantly, they refer to a process that has been automated and therefore can have major consequences because of the scale at which they can can operate [**234**] and authority that it is granted [**282**]. Whether or not this algorithm is a series of if-else statements or a machine-learned model with all of its complexity adds additional subtleties to the problem, but does not change the potential impact.

The phrases "representativeness", "bias", and "fairness" are not synonyms but reflect this general question of how close to some desired state is data or the outputs of an algorithm or platform. That is, are the benefits of these technologies evenly distributed or reflecting structural inequalities that exist in the world. Bias and fairness in particular can refer to many different aspects [**171, 296**]. For instance, in Chapter 8, the term "locality bias" is used to refer to the degree to which a place recommendation algorithm encodes a restaurant's location. This term assumes that a place recommendation model might encode information solely about aspects of the restaurants like cuisine or ambiance without learning any information about the restaurant's location. Thus, to the degree that the algorithm encodes location, it is biased. We examine the degree to which the outputs of place recommendation models are uniformly distributed across neighborhoods, with deviation from this uniform distribution being a measure of how "(un)fair" the algorithm's outputs are

(also known as algorithmic bias). And finally, "representativeness" or "equal representation" would ask whether the benefits of place recommendation algorithms are reflective of the amount of effort that a given community put into the platform. When I use these terms, I seek to define this proposed state against which I am comparing the data or algorithmic output.

2.2. Geography of Inequality

This dissertation focus on inequalities in online data and algorithmic systems, but these systematic disparities in representation have arisen, in part, due to a long history of offline discrimination and unequal access to resources. While any summarization of this long history will be incredibly reductive, it would be wrong to ignore this aspect. A few pertinent general points are discussed below then that pertain to the spatial nature of these inequalities.

If society was well-integrated and evenly-distributed with respect to race, class, and other demographics, then there would be no connection between physical location and inequality. An individuals zip code would carry no more information than where to route their mail. We might expect that all areas of the country would have plenty of content on Wikipedia and users on Foursquare.

Society is not spatially well-integrated though. Strong geographic patterns have arisen due to a history of residential segregation compounded by further self-sorting via spatial homophily. In the United States, the government enforced residential segregation through a wide variety of mechanisms—e.g., programs that provided insurance solely for housing developments that were white-only, preservation of residential segregation in public housing, allowing racially-restrictive housing covenants, a lack of protection from the police from violence and scare tactics intended to keep black homeowners out of white housing areas,

and the building of physical barriers that separated white and black communities. Within Chicago alone, there is substantial evidence of these policies: segregation and depressed housing values today can be attributed in part to federal redlining maps from the 1930s [**2**], the Chicago Housing Authority illegally reinforced segregation through the building of public housing solely in minority neighborhoods into the 1970s [**262**], the path followed by the Dan Ryan Expressway appears to have been relocated to serve as a physical barrier separating the predominantly white neighborhood of Bridgeport (home to the then mayor's family) and predominantly black neighborhoods to the east [**290**], racial covenants prevented black individuals from purchasing homes, famously in Hyde Park and Washington Park with the support of the University of Chicago [**95**], and rioters were allowed to attack homes and individuals perceived to be bringing black individuals into white neighborhoods [**131**].

Beyond cities, inequalities and differences have also arisen along the urban-rural divide (a common lens of analysis in this work). While urban and rural refer strictly to population density, the "divide" is an acknowledgement of the many differences between these two regions—e.g., cultural [**61**], political [**14**], access to healthcare [**94**], adoption of technology [**240**].

We examine differences in representation along the urban-rural divide (Part I) and with respect to race and income (Part II) in part because these populations are spatially distinct (and so easier to study through the lens of geography) but also because of these differences in culture and access to resources, which provide insight as to whether internet technologies are reducing inequalities as originally hoped [**153**] or amplifying them [**89**].

2.3. Online Representation

A core concept in this thesis is that of online representation, which I center around the idea that equal effort should lead to equal benefits. For example, if a certain proportion of a rural community edits Wikipedia, they should be able to achieve a level of quality of content that is equal to that of any other population who participates at the same rate. Or said another way, it should not be systematically more difficult for a given population to receive a given level of benefits from a given technology. This concept of "representation" is difficult to quantify when it comes to online platforms though. For comparison, in a political sphere, fair or equal representation might be quantified as each individual having an equally-weighted vote in an election [**67**]. What constitutes a vote or equal weight in an online setting is less clear.

Participation or coverage, as discussed below, are certainly important facets of representation online, but as I show in the first part of this thesis, there are many other factors that affect whether an online platform is representative. Measuring bias in representation in terms of unequal participation rates or coverage makes the implicit assumption that proportionate participation or coverage will lead to equal benefits and is the desired end-state. However, representation is more complicated than that in the online sphere. An area of the world could have no Wikipedia editors and yet outsiders could cull together sources and write very high-quality articles about the region—e.g., see [**274**] for some of these patterns. As I show in Section 5, having training data proportional to population is not always sufficient for equal algorithmic performance. On the other hand, Black Twitter was able to achieve a lot of visibility, despite black users being a minority on Twitter, due to their successful use of hashtags and resultant visibility via trending topics [**31**].

Together, these examples highlight the importance of not just measuring quantities (e.g., participation rates or coverage) but understanding the value of the content in the context of who or what is consuming it—i.e. does a population have a voice through their aggregate content in the context of users of the system or research and algorithms that learn from the data? These are not straightforward metrics to calculate, but the emphasis on not just the contributions but what is done with those contributions hopefully brings us closer to a more actionable understanding of online representation.

I consider three main consumers of user-generated content across my research—users (§4, research (§3, and algorithms (§5—but focus mainly on algorithms and research, as described below.

2.3.1. Participation in Online Platforms

While critiques of the appropriateness of studying human behavior or building algorithms from online data have focused on many challenges [**30, 265, 307**], studies that seek to quantify biases in this data often focus on participation and coverage. Surveys have helped to establish that different demographics—e.g., by race, age, education, income, population density—participate in social media [**130, 39**] and peer production [**144, 293**] at systematically different rates. Analyses of the content that exists on these platforms has also consistently shown population biases in participation and coverage biases in content (e.g., [**117, 140, 181, 187, 207, 222**]).

There is evidence that certain structural barriers to participation are lowering, but that plenty of obstacles remain that hinder equality of participation and coverage in online platforms. While gaps in access—e.g., broadband access in the United States [**39**]—have decreased in many areas, Graham et al. [**112**] have further shown that broadband access is a

necessary but not sufficient condition for participation on Wikipedia. The degree to which platform norms or design discourage use by certain populations has shown up in many other contexts—e.g., women on OpenStreetMap [**293**], Wikipedia [**217**], and mobile crowdsourcing platform Taskrabbit [**301**], rural users and location-based social networks [**129**]. There is some evidence that even without an explicit racial structure to the internet, offline patterns of segregation have arisen—e.g., [**213**].

2.3.2. UGC and Research

The field of computational social science has explored how UGC (predominantly social media) might be used as a means of conducting observational studies of people. As noted in *Science* [**265**] and the subject of workshops (e.g., [**7**]), this lens has been quite powerful when it comes to the study of many of these processes. The common procedure taken by these studies is to first filter, generally via heuristics, social media posts into buckets that are associated with specific populations—e.g., geographically filter Flickr photos based on geotags to those associated with residents of a given city, demographically filter tweets for those from college students based on self-identification in a profile, topically filter posts based on hashtags for those associated with a given social movement. These posts are then analyzed so as to reach conclusions about that general population, such as the happiness of a given population or common trajectory of a movement. This approach has been taken by researchers in HCI (e.g., [**3, 56, 306, 338, 178**]), the social sciences (e.g., [**58, 97, 146, 271, 308, 348**], and even the natural sciences (e.g., [**298, 324, 266**]. They have explored phenomena of interest ranging from social unrest and emergencies [**51, 175**] to disease tracking [**147, 182, 284**].

Importantly, if these computational social science studies are to reach the correct conclusions about not just people who use a platform in a given manner that matches their

heuristics, but a population of people more generally, then they must account for biases in this content. This challenge has been raised by many (e.g., [**265, 307, 57**]), but detailing where UGC fails in this regard and how to adapt methods for this is less common. The research detailed in Chapter 3 seeks to provide insight into how well UGC represents rural populations, specifically by examining what proportion of content in these areas is "local" and how that impacts research conclusions about those areas.

2.3.3. UGC and Algorithms

User-generated content has also served as a rich source of training data for algorithms. Even without crowdsourcing labels, this data underlies many AI technologies—e.g., conversational agents (Reddit and Alexa [**71**]), entity disambiguation (Wikipedia and OpenAI [**253**]), knowledge graphs (Wikipedia and Google [**289**]), word embeddings (Wikipedia, Twitter, and Common Crawl in GloVe embeddings [**238**]), object detection (Instagram hashtags and Facebook [**292**])—and adding supervision has allowed for training of even more specific computational models—e.g., harassment detection (Wikipedia comments [**325**]), image segmentation and captioning (Flickr photos [**195**]), discussion labeling (Reddit threads [**339**]), reading comprehension (Wikipedia articles [**254**]).

Research has established that coverage biases are tied to algorithmic bias for models that are trained on this data. Still lacking though is a deeper understanding of which types of algorithmic biases can be corrected via greater coverage alone and which types require interventions such as different types of algorithms.

For instance, Culotta and colleagues found that adjusting for who was contributing content on Twitter improved performance in predicting public health metrics [**57**] and text-based geolocation [**183**]. In contrast, Pavalanathan and Eisenstein [**237**] did not see improvements

to text-based geolocation when adjusting for population imbalances in Twitter data. They found that algorithmic bias along age and gender lines was not driven by undercoverage. Buolamwini and Gebru [**34**] found algorithmic bias in facial analysis algorithms with respect to skin color and gender and that more balanced training data largely fixed the issue. Blodgett et al. [**27**] found that language identification algorithms consistently misidentify tweets from African-American individuals at higher rates than tweets from White individuals and built an ensemble classifier that incorporates additional demographic inference to address these disparities.

There are many more examples of these algorithmic biases that make it abundantly clear that there is no one, simple fix. Greater transparency and an understanding of the mechanisms by which algorithmic bias arise and might be addressed are important to moving forward, though just one part of achieving more accountable algorithms [**62**]). The second part of this dissertation focuses on how we might address biases within geographic algorithms, especially those that are not addressed through simple increases in coverage. I do not examine the design of online platforms or social interventions to elicit more representative content—e.g., NextDoor [**142**], Uber [**260**], Wikipedia [**119**]—which aim for long-term (and incredibly important) fixes to these challenges. Instead, I consider how to improve our ability to evaluate existing algorithms as well as transformations or algorithmic designs that can be applied given the current, biased nature of this data. This is not undertaken with the belief that fixing the algorithm will fix the problem (or even improve the situation) but with the desire to start a more robust conversation around the design of these algorithms and choices that affect their outcomes as well as identification of more human-centered metrics that capture some of the impact of these technologies on communities and might be used to evaluate these algorithms.

2.4. Geographic Algorithms

The second half of this thesis focuses specifically on the domain of geographic algorithms and how they might be designed more responsibly given the challenges surfaced in the first half of the dissertation. In the context of preventing "technological redlining" [**232**], we consider a geographic algorithm to be one that might have systematically differential performance (and impacts) across different geographic regions. Based on this criteria, a few algorithmic technologies stand out due to their ubiquity and potential for impact: vehicle routing (e.g., Google Maps driving directions), place recommendation (e.g., Yelp), location-based games (e.g., Pokémon GO), predictive policing (e.g., PredPol). Other classes of algorithms certainly have geographic components and would warrant further study—e.g., many language models have geographic aspects to them [**74**] and can see deteriorated performance for certain populations [**27**], there is increasing recognition that geography matters for objection detection algorithms [**66**], there are many geographic biases associated with the sharing economy [**301**] though the connections to the algorithmic components of those platforms is less well studied. The final chapter in this dissertation (§9) takes a step towards more general evaluations, in the realm of geographic representation learning as a core component of many classification algorithms.

2.4.1. Impact of Geographic Algorithms

A major motivation for focusing on geographic algorithms is that many of the platforms they support have large social and economic implications. For example, Pokémon Go reached more daily users than Twitter [**81**] and many of these users traveled to new neighborhoods and businesses in search of in-game elements while playing [**48**]. Yelp has close to 100

million monthly active users [**1**] and Luca [**198**] has demonstrated that a one-star increase in a restaurant's Yelp ratings could lead to 5-9% more revenue for that restaurant. Predictive policing algorithms have been shown to be susceptible to runaway feedback loops that can distort the true geographic distribution of crimes [**80**]. Pedestrian and vehicle traffic is linked to restaurant revenue [**344**] and algorithmic adjustments made by Waze have had quite salient impacts on neighborhoods in terms of traffic [**337, 208**].

This social and economic importance of geographic algorithms motivates the need to evaluate whether they are also fair. In line with the well-documented "spatial is special" [**141**], however, evaluating geographic algorithms for whether they are fair is often not straightforward. There are at least two pertinent challenges that make this undertaking substantially different from much of the adjacent fairness literature (e.g., [**72, 5, 54**]): the consumer is not the key constituent with regards to fairness and the outcomes are neither binary nor directly tied to race or other sensitive attributes.

2.4.2. Fairness of Geographic Algorithms

Many algorithms considered in the fairness literature are evaluated from the consumer perspective (C-fairness per [**37**]) and have binary outcomes that can then be explicitly tied to race or other attributes—e.g., are black and white individuals treated equally by algorithms that determine recidivism risk [**54, 171**] or school admissions [**180, 5**]. This focus on the consumer means that if you have a representative group of individuals from each group—e.g., where race determines group—then you can evaluate factors such as false-positive rate or calibration to determine whether the algorithm is "fair". While there are challenges in

how to define fairness and the inherent tension between fairness metrics [**171, 54**], the calculation of these statistics for each sensitive attribute once a metric is chosen is generally straightforward.

The impact of the geographic algorithms like place recommendation or vehicle routing, however, is most salient in terms of the items being recommended—e.g., does a street receive traffic, is a restaurant recommended to diners, does a neighborhood see an increased police presence? This focus on the items, as opposed to the consumer, is known as provider fairness or P-fairness [**37**]. The goal is to ensure that the benefits of being recommended are distributed equitably. Because fairness in this case is calculated as a function of all of the outputs, and not just e.g., the false-positive rate for a given group, the inputs themselves must be representative of the relative distribution. That is, if you want to make a statement about whether vehicle routing algorithms are unfairly favoring certain types of neighborhoods, you have to have a good idea of where people start and end their routes so you can determine where in the city the algorithm might have an impact. Public datasets that provide a good proxy for the distribution of these inputs can be difficult to impossible to attain for geographic algorithms because this data is often highly sensitive at the granularity that is needed. Additionally, because the items being recommended are not directly tied to sensitive attributes but instead are often regions, additional choices have to be made about how to connect an entity like a street or restaurant to key attributes such as race.

2.5. Geographic in Computational Models

The final study of this dissertation (§9) and to some degree the evaluation of place recommendation (§8) are examinations of geographic representation learning, a core component

of many geographic algorithms. The following related work is intended to provide a framework and motivation for these studies, as well as an indication of how well they and other studies of bias in user-generated content might generalize (or require complementary studies to study algorithms that have encoded different choices). It is a survey of the many ways in which geography is represented within statistical models. It covers four core aspects of this representation: how do we represent a given spatial data point, how do we determine the relatedness between two points, how do we aggregate points, and how do we evaluate the resulting model. I refer to the following as *geographic hyperparameters*—i.e. choices that are made in statistical models of spatial processes—to make clear that even though some have very established default practices, choosing the default is still a choice with implications. Notably, whereas for situations such as laying the infrastructure for internet connectivity, we do not have the choice of defining distance by anything other than Euclidean distance, in the design of algorithms or research, we do have alternative choices that might often be more equitable and effective.

2.5.1. Representing a Point

Consider a data point that corresponds to a location in the physical world (e.g., a city block and how safe that location is perceived to be). In order for a predictive model to effectively generalize what it learns about that data point to other potential locations, it must represent geography in such a way that similar places have similar representations. A model can then calculate the similarity (or distance) between two given locations (and therefore how pertinent this data point is predictions made for those other locations).

2.5.1.1. Physical Location. Many approaches to determining the similarity between two locations rely on physical distance, thereby operationalizing the literal interpretation of

Tobler's First Law [**303**] and accounting for what is often called distance decay [**301**]. This approach represents geography as a two-dimensional vector—i.e. one dimension for latitude and one dimension for longitude.

2.5.1.2. Regions. There is a long history to regional geography and types of regions such as higher-order administrative units are known to be important for understanding many processes in human geography [**25**]. To capture these similarities, a given point can be said to be contained within several higher-order administrative units (e.g., cities, counties, states, countries). To reduce computational complexity, various grid systems have also been introduced to serve as replacement regions—e.g., S2 cells [**321**] or researcher-defined grids [**251**]. Little attempt is made to align these grids with existing boundaries, but, with a suitably small grid-size, misalignment between the grid and administratively-defined regions can be made minimal.

This effectively operationalizes geography as a one-hot encoding in which a given location can be one of many classes. The one-hot encoding is sparse though and misses the many dependencies between spatial data that span borders. This independence means that regions alone also provide limited predictive value within models in that they do not help with extrapolating data to new regions.

2.5.1.3. Attributes. A further step in enriching a geographic model is the incorporation of attributes about a given area. For example, this may take the form of structured census data as a proxy for place (e.g. household median income, race, how urban an area is). This can extend the dimensionality of a given place's representation from a few dimensions to almost arbitrarily more (i.e. an additional dimension for each census variable). This helps a model to understand similarities between areas that are not monotonic with distance and that arise from more complicated interactions of characteristics that describe a place. For

instance, FiveThirtyEight's election models incorporate race and religion variables [**287**], Clewlow and Mishra [**47**] incorporate gender, race, age, education, income, and population density into modeling of ride-share platform adoption, and Culotta [**56**] explores the value of race, age, gender, and income (as well as signals from Twitter) in predicting the health indicators for a given county.

While powerful, census variables (or other population metrics) also bring substantial constraints though that limit these models. Census data is not generally consistent in how it is collected or represented across different countries, making it difficult to include census data in geographic prediction models that span multiple countries. The data is also generally aggregated to specific predetermined administrative areas (e.g., census tracts, counties), which leads to substantial ecological validity challenges if these administrative units are not an appropriate scale for studying the problem at hand. Finally, joining the census data with data not collected at the same scale can be computationally-intensive (e.g. point-in-polygon operations).

2.5.1.4. Trace Data. More recently, in order to overcome some of the limitations mentioned above of traditional geographic representations, researchers have begun to explore how to incorporate geographic user-generated content and sensor data into geographic modeling and prediction algorithms. Some of these approaches are listed below and cover a wide variety of datasets and tasks. Researchers have built models to predict population demographics through image processing techniques applied to Google Street View imagery [**100**], satellite imagery [**152, 297**], photos in geotagged tweets [**4**], aggregate check-in behavior on Foursquare along with business information from OpenStreetMap [**309**], and linguistic analysis of geotagged tweets [**56**]. Prediction of the perception and character of places (e.g., how safe a place feels) has been done through Flickr tags [**162, 251, 252**], Google Street View

Table 2.1. Geographic data sources.

Framework for categorizing geographic data by type of media and type of source.

Data Format	Possible Representation Learning Method	Data Sources (i.e. what is being encoding)		
		Social Media	Peer Production	Sensor / Collected Data
Point sequences	Skip-gram modeling (e.g. word2vec), matrix factorization of co-occurrence data	LBSN traces (e.g. Foursquare), consecutive geotagged contributions (e.g. locations of Yelp reviews or Flickr photos taken by a given user)	OpenStreetMap geometries (e.g. roads), links between Wikipedia articles	Mobile phone traces, migration data, Wikipedia navigation data (e.g. consecutive pageviews by a given user to geographic articles)
Text	Paragraph-vector (e.g. doc2vec), topic modeling (e.g., LDA)	Geolocated microblogs (e.g. Twitter), photo tags/comments (e.g. Flickr)	OpenStreetMap tags (e.g. amenity:restaurant), Wikipedia articles (e.g. about the Washington Monument)	Survey data, text inputs (e.g. search query logs)
Imagery	Histogram of Oriented Gradients (HOG), Pre-trained CNN	Photos (e.g. Flickr, Instagram)	Mapillary street view imagery, Wikimedia Commons	Satellite imagery, Google Street View imagery

imagery [**69, 226**], OpenStreetMap places [**212**], and Foursquare images, ratings, and check-ins [**55, 185, 270, 340**]. Geographic similarity algorithms have been learned from Wikipedia articles [**272**] and migration patterns modeled through Flickr photo sequences [**22**].

As illustrated by this long list of papers that have used online trace data in geographic models, there are many possible ways of representing a location through media associated with that place. Different types of media (e.g., text, images) require different representation learning algorithms to make them compatible with an algorithmic system, and different sources of geographic data (e.g., social media, peer production) emphasize different types of signals and therefore what is encoded. Table 2.1 provides a summary and examples of some of these options below.

The data sources fall on a spectrum with regard to whether they represent more individual experiences of place or objective recordings of what exists there. For example, geographic social media (e.g. tweets) are quite noisy, unevenly representative, and unstructured but provide potentially rich insight into the nature of different places while sensor data (e.g. satellite

imagery) can be much more structured and representative but potentially far more limited in reflecting how individuals use and feel about the space. While there are exceptions (e.g., Facebook has a very broad user base), social media generally has the greatest coverage biases while sensor data the least ([**136, 140**] as well as §4). Point sequences are generally the simplest data representation and a major form of trace data, making them more representative and easier to process. As I show in Chapter 8, they also can encode strong biases around location though. Text and imagery can be quite complex due to a shifting (visual) language, but also potentially can bridge biases from segregation that human mobility patterns encode.

2.5.2. Relatedness of Two Points

For a model to generalize what it learns about one place to another, it must have some way of relating different places that is independent of what is being measured. Within a geographic prediction model, physical distance has generally been used to extrapolate data to new areas—e.g. the value for a new location can be estimated as the weighted average of the values from nearby locations. The exact approaches to converting distance to similarity vary greatly though: continuous function of distance [**84, 230, 247, 328, 331**], step function where only locations within a certain range are considered to be similar [**10, 33, 83**], ordinal distance [**233, 272**] that measures the number of entities such as POIs between two locations to better capture variation in population density. While distance does effectively capture many similarities between places, it only takes advantage of two dimensions and becomes much less useful across much larger regions—e.g., it provides very little information about the relative similarity of cities that are thousands of miles away from each other [**190**].

Each of the other point representations mentioned above—regions, attributes, trace data—has corresponding methods for computing distance as well. Regions might use containment: are two regions, such as cities, contained within an even larger region, such as a state. Models that use attributes or trace data generally rely on general metrics that measure the similarity of vectors like Jaccard similarity or cosine similarity. The challenge for these models is how to weight or represent the attributes or trace data such that the most important components are what determines similarity.

2.5.3. Aggregating Points

A closely-related decision to the choice of how to operationalize distance is the scale and shape at which the data will be represented—i.e. as points or aggregated to areas such as grid cells or administratively-defined units like states or countries. The choice of scale can determine the conclusions one reaches about a given region, which is known as the Modifiable Areal Unit Problem [**87**]. For instance, spatial processes such as segregation often occur at the scale of neighborhoods, which second-order units such as counties would fail to capture. Chicago can be viewed as both one of the most integrated and most segregated cities in the United States depending on whether you calculate segregation at the neighborhood- or city-scale [**286**]. To ensure robustness, spatial modeling is occasionally done at multiple scales—e.g. studying political polarization in the United States at the level of census divisions, states, and counties [**159**]—but the availability of data and complexity of spatial modeling often precludes this from occurring. Some of the choices for aggregation are described below.

2.5.3.1. Administrative Units. Administrative units are designed by governments, generally to capture regions with apparent boundaries and maximum demographic homogeneity[1]. While administrative units often exist at many scales (e.g., census tracts, zip codes, cities, counties, states, countries), it can be computationally- and human-intensive to gather all of these boundaries and join them to the data. Openly-available administrative boundaries (e.g., GADM's database) are often not available any finer than "second-level" units for most countries. This is the equivalent of counties in the United States, which contain anywhere from 88 people in Kalawao County, Hawaii to over 10 million people in Los Angeles County, California. Furthermore, it can be very computationally intensive to compute the point-in-polygon operations necessary to link a given data point (e.g. image taken at a given location) with all of the regions that it is determined to be contained within (e.g., census tract, city, state).

Table 2.2 details various common choices for administrative units and their distribution of populations and areas, and Table 2.3 shows how these different choices would change the representation of a specific location in Chicago. Notably, certain units have much more consistent population or area, so, depending on the process that is being modeled, different choices might impose differential burdens of generating data on regions. Grid cells often have an approximately fixed area, though aggregating them smartly can create variation that leads to consistent population counts.

2.5.3.2. Grid Systems. Grid systems offer a more practical and scalable alternative to administrative boundaries as the unit of aggregation, but their effectiveness at capturing human geography is less clear. Grid systems generally are simple to represent, rely on

[1]https://www2.census.gov/geo/pdfs/reference/GARM/Ch10GARM.pdf

Table 2.2. Descriptive statistics for various administrative spatial units.

Common choices of administrative units in the United States and their number of units, range of area, and range of population.

Administrative Area	Total Units	Population Range	Area Range (kmˆ2)
States	50 + 1 (DC)	580,00 - 39,000,000	177 - 1,700,000
Counties	3,144	88 - 10,100,000	31 - 52,000
Census Tracts	73,056	1,200 - 8,000	0.02 - 234,600
Zip Code Tabulation Areas (ZCTAs)	32,989	0 - 115,104	0.005 - 34,900
Congressional Districts	435	528,000 - 995,000	26.55 - 1,481,350

Table 2.3. Administrative units in Chicago.

The range of ways that a specific point—the Children's Museum of Art and Social Justice, located at 2007 S Halsted St, Chicago, IL 60608, or, alternatively the latitude/longitude point of 41.8550, -87.6464—can be represented as part of a broader administrative unit. The percent of the population that is non-white in each administrative unit is also given to show how that variable changes based on aggregation.

Name	Area (sq. miles)	Population	Percent Non-White
State of Illinois	55,517	12,801,539	38%
Chicago-Naperville-Elgin, IL-IN-WI Core Based Statistical Area	7,197	9,512,968	47%
Cook County	945	5,203,499	58%
City of Chicago	227	2,704,965	67%
Congressional District 4	52	737,025	42%
Zip Code 60608	6.3	78,072	83%
Census Tract 3102	0.1	1,521	66%
Census Block Group 2	0.049	686	67%
Census Block 2021	0.003	10	10%

hashing or other constant-time functions to determine which cell contains a given latitude-longitude point, and have natural hierarchies (i.e. nested scales such that any given cell can be evenly divided into a predetermined number of smaller cells). For instance, Weyand et al. [**321**] used S2 cells for flexibly geolocating photos around the world, and Quercia et al. [**251**] used a 200m-by-200m grid for consistently computing the beauty of a given location from Flickr photo tags in the cities of Boston (USA) and London (UK).

2.5.4. Model Evaluation

Finally, the choice of geographic context for validating a given model can greatly impact its performance, fairness, and what the model learns is important. For instance, researchers have also documented decreased precision predicting human perceptions of urban landscapes

and object detection in a city when images and labels from a more distant city were used to train the model [**226, 66**], different levels of anonymity in location-based social networks in different regions [**261**], and varying effectiveness of the sharing economy depending on the socioeconomic status of a given area [**301**]. The Inclusive Images Challenge by Google [**66**] documents this challenge with regard to image object detection. Together, these examples highlight the importance of geographic context for evaluating models, something that is explored in more depth in Chapters 7–9.

CHAPTER 3

Localness and Social Media

In this chapter, we consider how "local" various forms of social media content are in urban and rural areas. We look at the impact of non-local content on the conclusions of observational research that compares the characteristics of these regions.[1]

3.1. Introduction

Social media volunteered geographic information (VGI) such as geotagged tweets, geotagged photos, and "check-ins" provides an unprecedented real-time lens into many important spatiotemporal processes. As detailed in Section 2.3.2, there is a large body of HCI research that has conducted observational studies using social media. A common thread in these studies of social media VGI is the reliance on a simple assumption we call the *Localness Assumption*. Under this assumption, which is almost always adopted implicitly, *a unit of social media VGI always represents the perspective or experience of a person who is local to the region of the corresponding geotag*. Put more simply, the localness assumption presumes that social media users can be considered locals from everywhere they post geotagged content. For example, adopting the localness assumption, one can assume that a person who posts a geotagged tweet about a political candidate is doing so from her or his home voting district, thereby affording applications like election forecasting and political preference monitoring.

[1]The work presented in this chapter was originally published in: **Johnson, I.**, Sengupta, S., Schning, J., and Hecht, B. The Geography and Importance of Localness in Geotagged Social Media. *ACM Conference on Human Factors in Computing Systems 2016*.

Human mobility is the major potential confound to the localness assumption. Studies that adopt the localness assumption implicitly argue that people who post geotagged social media while on business trips, vacations, and other forms of travel are not a significant factor across large datasets of social media VGI. This consideration of human mobility and the localness assumption more generally dates back to the origins of the term "volunteered geographic information": in the foundational paper on VGI, well-known geographer Michael Goodchild argued that the core value of VGI is that it tends to come from locals [**109**]. However, Goodchild was writing in a time largely before smartphones and social media, a time when it might be reasonable to assume that human mobility may be dampened in VGI datasets.

In this chapter, we present the first systematic examination of the validity of the localness assumption in social media VGI. Analyzing four datasets across three distinct types of social media VGI, we find that, due to human mobility, *the localness assumption does not hold for approximately 25% of social media VGI.* Additionally, we identify extensive geographic variation in the localness of social media VGI. In other words, while Goodchild's "localness ideal" – in which VGI contains high-quality local information – holds somewhat true in certain areas, there are other areas where the connection between the population contributing social media VGI and local population is much more tenuous. Of particular concern, we find that the degree of localness in an area tends to compound previously identified population biases in social media VGI [**140, 187, 222**], with rural and older areas not only having a diminished voice overall in social media VGI, but (as our results show) that voice is diluted by outsiders at a disproportionate rate.

Through a case study focusing on recent work that assesses the "geography of happiness" in the U.S. through geotagged tweets [**223**], we explore the direct effect of the localness

assumption on social media VGI-based studies. We replicated the approach employed in ([**223**]) and compared its output to that of several versions of the approach that explicitly filter out non-local tweets (thereby accounting for human mobility and not adopting the localness assumption). We found this filtering process resulted in small shifts in the happiness geography of the United States overall, but that there were significant shifts in certain key types of regions, highlighting the importance of filtering out non-local content when doing social media VGI-based research.

This chapter also makes an important methodological contribution: this chapter is the first to aggregate and compare the various methods for determining VGI localness in the small literature that has explicitly considered localness, finding important differences between these methods. We discuss the implications of these results and outline best practices for filtering out non-local content in studies that use social media VGI.

To summarize, this chapter makes the following contributions:

- We characterize the amount of non-local content in multiple social media VGI repositories, finding that approximately 25% of geotagged content on average is non-local to an area.
- We find that the geographic contours of localness follow important sociodemographic properties, with, for example, more urban and younger areas having consistently greater proportions of local content.
- We examine the impact of non-local VGI in one of the many studies that adopts the localness assumption. We show that the presence of non-local VGI can significantly alter algorithmic determination of regional properties (e.g., "happiness" retrieved from tweets) for certain areas.

- We also make an important methodological contribution: we characterize the various definitions of localness in use by the research community, find clear differences in their results, and outline a series of corresponding best practices for social media VGI studies.

Below, we first cover related work and detail the social media VGI datasets we consider. We then present our methods and results, with this discussion structured around four research questions. Each of these questions corresponds to one of the contributions listed above. We close by highlighting the broader implications of our work, a discussion that includes a series of localness best practices for social media VGI researchers.

3.2. Related Work

This research is primarily motivated by three areas of prior work: (1) research on localness in VGI, (2) research on population biases in social media, and (3) studies that use social media VGI.

3.2.1. Localness in VGI

Key inspiration for this work came from the work of Sen et al. [**274**] that demonstrated that, at a country-to-country scale, there is extensive geographic variation in the localness of geographic content in Wikipedia (peer production VGI) and that this variation corresponds to global socioeconomic contours. This work can very broadly be thought of as an extension of Sen et al.'s work to the social media VGI domain. Social media VGI has a fundamentally different "spatial content production model" [**136**] than peer production VGI, which suggests that its localness dynamics will be substantially different (i.e. to post a geotagged tweet, one has to be at the location of the geotag, but one can write a Wikipedia article about

anywhere in the world from anywhere in the world). This work also addresses the call in Sen et al. for localness work that considers localness at a spatial scale more granular than that of the country (we study localness at the U.S. county scale).

Sen et al. is not the only work to consider localness in a peer production VGI context. Other, more peripherally related research includes work establishing that local peer-produced VGI is of higher quality than non-local contributions [**73, 354**], modeling Wikipedia contributions as a spatial process [**128**], and examining global core/periphery dynamics between Wikipedia editors and the geographic articles that they edit [**113**].

Additional core motivation for this work is derived from Hecht and Gergle [**136**], which examined the distance between content contributors and the subjects of their contributions in Wikipedia and Flickr. To our knowledge, this is the only other research to directly consider localness in a social media VGI context, finding that while the computer vision community had often assumed that that Flickr photos primarily came from tourists (the opposite of the localness assumption that is pervasive in social media VGI research), many Flickr photos come from locals (a dynamic beautifully visualized in the maps created by Fischer [**86**]). This research builds on that of Hecht and Gergle by directly targeting the localness assumption that is central to so many studies that utilize social media VGI, characterizing the degree of localness across three separate types of social media VGI, its geographic and sociodemographic variation, and its effects on studies. Moreover, this is the first work to problematize the operationalizations of localness that have appeared in the literature (including that in the work of Hecht and Gergle [**136**]) and identify corresponding best practices.

3.2.2. Population Bias in Social Media

Section 2.3.1 provides a high-level overview of population bias in social media, but a few examples are useful for framing this work. For instance, Li et al. [**187**] found that the density of tweets per capita and geotagged Flickr photos per capita are positively correlated with sociodemographic factors like income, youth, and education (when comparing county-aggregated values in California). Similar work with Twitter was done by Malik et al. [**201**], but across the United States. Hecht and Stephens [**140**] focused specifically on the rural/urban divide, finding that in Foursquare, Twitter, and Flickr, there was far more content per capita in urban areas than rural areas. As we will see below, when it comes to geographic variations in localness, the same types of areas that tend to be advantaged when it comes to raw quantity of social media VGI (e.g., urban areas) also tend to be advantaged in terms of the their VGI's localness.

3.2.3. Social Media VGI Based Studies

Section 2.3.2 provides an overview of social media VGI-based studies. Below, we show through a case study on the work of Mitchell et al. [**223**] what can occur if non-local VGI is filtered out of these studies. Notably, localness likely does not pertain to most studies that use geotagged social media for sensing natural events such as earthquakes or disasters (e.g., [**175, 266**]).

Several studies have made intentional efforts to filter out non-local tweets, thus explicitly eschewing the localness assumption. For instance, the work of Li et al. [**187**] and Hecht and Stephens [**140**] fall into this class of research. However, as we will see below, these studies and others that have separated local and non-local tweets take unique approaches to doing

so, with each approach operationalizing very different understandings of what it means to be a local.

Finally, it is important to note that a much larger body of work unintentionally works around the localness assumption by using data from the location fields of social media users' profiles rather than geotags [**21, 225, 177**]. While this data source has serious problems that geotags do not (e.g., [**138, 314**]), when valid, associating a user's social media with their self-described home location rather than the locations of its geotags is one technique to filter out non-local social media. As such, when comparing approaches to quantifying localness in social media VGI, we consider the use of the location field as one such technique.

3.3. Datasets

In order to gain a broad, ecologically valid understanding of localness phenomena in social media VGI, we look at VGI from three different types of popular social media communities: a microblogging platform (Twitter), a photo-sharing community (Flickr), and a check-in based location-based social network (Swarm). In this section, we describe our data from each of these communities (and other sources) in more detail.

3.3.1. Twitter

We analyze two datasets of geotagged tweets: both were gathered through the public Streaming API and were restricted only to tweets with geotags (latitude and longitude coordinates). The first dataset, which we shall refer to as *T-51M*, was gathered from October 19, 2014, through November 19, 2014. It contains 51.2 million tweets in the contiguous United States from 1.6 million users. The second dataset, *T-11M*, was gathered from May 27, 2015 through August 19, 2015. It contains 10.8 million tweets from 964,000 users in the contiguous United

States[2]. We combine these two datasets for our study of happiness to improve robustness and incorporate data from different times of the year. The combined dataset has 61.9 million tweets from 2.2 million unique users. For all Twitter datasets, we remove organizational accounts per best practices for social media research [**265**] prior to analysis through use of the classifier described by McCorriston et al. [**209**]. Organizational tweets comprised 6.3% of the total dataset.

3.3.2. Flickr

For Flickr, we analyze the YFCC100M dataset, which we shall refer to as *F-15M*. It contains 15.4 million geotagged Creative-Commons-Licensed photos from 73,797 thousand users in the contiguous United States. The YFCC100M dataset [**302**] was publicly released by Yahoo Labs and Flickr in June 2014.

3.3.3. Swarm

Swarm (formerly Foursquare) is a location-based social network in which users check in to locations and broadcast this information to their social network. The Swarm API does not allow public access to check-in data, but some users choose to publicly tweet their check-ins. Swarm check-ins shared via Twitter have been used extensively in the past (e.g. [**96, 233**]). We analyze the dataset collected by Cheng et al. [**45**] from September 2010 through January 2011 consisting of 7.8 million check-ins from 89 thousand users for the United States, which we shall call *S-8M*.

[2]The size difference in these datasets arises from using a contiguous United States and global bounding box respectively.

3.3.4. Sociodemographic Statistics

Several of our analyses involve comparing sociodemographic information to the percentage of VGI in a U.S. county that is local (U.S. counties are second-order administrative districts, right below states). All of the sociodemographic variables we examine relate to population biases that have been detected in prior work with social media VGI: from the 2010 US Census, we examine urban/rural (*% Urban)* [**140, 201, 222**], race (% White, Non-Latino or *%WNL*) [**187, 222**], and age (median age or *MedAge*) [**187, 201**]. From the 2009-2013 American Communities Survey, we examine income (household median income or *HMI*) [**187, 201**] and % Management, Business, Science, and Art occupations (*% MBSA*) [**187**].

Because these data are limited to the United States, we focus in this chapter on social media with geotags that fall within U.S. borders. Additionally, due to the requirements and assumptions of our spatial modeling approaches, analyses are limited to the contiguous United States (i.e. the "lower 48").

3.4. Research Questions

In this research, we posed four separate research questions, each of which corresponds to one of the contributions enumerated above. Specifically, we asked:

- **RQ0:** What precisely does it mean for a unit of social media VGI to be local to a given region?
- **RQ1:** What percent of social media VGI is local? In other words, to what extent is the localness assumption true for social media VGI?
- **RQ2:** What is the geography of localness within social media VGI? Does variation in localness follow the same socioeconomic contours that are seen in other types of volunteered geographic information?

- **RQ3:** How does the inclusion of non-local contributions impact the results of research and algorithms that leverage social media VGI?

In the following sections, we present our methods and results associated with each research question. This is followed by a holistic discussion of the implications of our results.

3.5. RQ0: What is Local?

3.5.1. Methods

To address the challenge of defining localness (and what it means for geotagged social media to be local), we turned to the limited social media VGI literature that has made efforts to separate local and non-local information. Conducting the first survey of techniques for distinguishing local from non-local content based on geotags, we identified four approaches in this literature, with each quantifying a different definition of localness. We implemented all four of these approaches, and use all four throughout this chapter.

In order to gain a better understanding of each localness approach and how it relates to the others, we classified every unit of social media in all four repositories as either local or non-local according to each approach and compared and contrasted the results. Below, we first describe the four localness approaches in more detail (as well as cover the key role that spatial scale plays in all of them). Following that, we discuss the results of our comparative analysis.

3.5.1.1. "n-days" Localness Metric. The *n-days* localness metric [**140, 187**] takes all of a user's contributions (e.g., tweets, photos, check-ins) and assigns the user as local to a given region (e.g., county, city) if they made contributions in that region at least n days apart. In order to be considered a local, this metric thus requires a person to demonstrate that they have either spent at least n days in a particular region or returned there at a later

date at least n days after they initially contributed. A sufficiently large choice of n must be used to filter out people who are just traveling through a region. The choice of n has varied, but both Hecht and Stephens [**140**] and Li et al. [**187**] use 10 days as the minimum length of time. We follow suit and set n equal to 10 days. Note that the *n-days* metric operationalizes an idea of localness in which a person can be local to between 0 and m regions, where m is the number of regions being studied – e.g., the number of counties in the U.S. – although in practice the number of local regions tends to be low.

3.5.1.2. "Plurality" Localness Metric. The *plurality* metric [**140, 224**] assigns a user as local to the region in which they contributed the most social media VGI in a given repository. Uniquely, this algorithm ensures that even users that do not make frequent contributions (e.g., who for example might be filtered out by the n-days algorithm for sheer lack of content) will still be included in the analysis. *Plurality* assigns each user as local to exactly one place, except in ties in which case all regions at that level of contribution are local.

3.5.1.3. "Geometric Median" Localness Metric. The *geometric median* metric [**161**] has been most commonly used in the geolocation inference literature to assign a home location to users. We implement the multivariate L1-median definition used, for example, by Jurgens et al. [**161**] and Compton et al. [**52**], which defines the median of a set of points as the point in space that minimizes the distance between it and all of the points in the set. We further require, per Jurgens et al. [**161**], that users have a minimum of five VGI points and that the median absolute deviation of the user's points to their geometric median be no greater than 30 km (i.e. half of the user's points must be within 30 km of the geometric median).

3.5.1.4. "Location Field" Localness Metric. The *location field* metric [**138, 161, 243**] has been used heavily for expanding social media VGI datasets beyond just explicitly geotagged social media, which often make up a small overall percentage of these datasets [**21, 177, 225**]. As noted above, this approach uses the self-reported location information in the "Location" field in users' social media profiles, which exists for all three social media communities considered here. The accuracy and completeness of location field data has been problematized by Hecht et al. [**138**], but its use continues as location field data is one of the only ways to geolocate the large percentages of social media users who do not geotag their content.

In order to turn a textual location into a machine-readable latitude/longitude coordinate, a *geocoder* is necessary. We used Jurgens et al.'s Geonames-based geocoder [**161**], which builds on the Creative-Commons-Licensed Geonames places dataset and handles noisy text through a series of regular expressions and common replacements (e.g., St. and Saint). We further validated the implementation by comparing our results to those achieved by use of Wikipedia redirects as implemented in the WikiBrain library [**273**] and described in [**138**]. For location field entries that both tools could geocode (˜54% of the Geonames results), there was 90% agreement.

3.5.1.5. Choosing the Correct Spatial Scale. A final source of variation in how localness has been operationalized in the literature occurs in the spatial scale of the localness definition. For example, Sen et al. [**274**] define a local Wikipedia editor as someone who edits an article about a place in the editor's home *country*, whereas Li et al. [**187**] define local at the U.S. *county* scale. In this chapter, we focus on the county-scale for two reasons: (1) it is a common scale at which social media VGI research is done (e.g., [**56, 140, 187**])

and (2) it is a scale at which the sociodemographic information we need to address RQ2 is available.

Because our definition of localness is at the U.S. county-scale, this means that the "regions" operated on by *n-days* and *plurality* are U.S. counties. For instance, if a person tweets predominately from places within Cobb County, GA, under *plurality*, all tweets from this person with geotags within Cobb County will be considered local, and those outside Cobb County will be considered non-local. Unlike *n-days* and *plurality*, the *geometric median* and *location field* metrics map users to a point. In these cases, we use simple point-in-polygon operations to assign the user as local to the county that contains the point.

3.5.1.6. Putting It All Together: Calculating Localness. To make the process of calculating localness more concrete, let us consider the case of a tweet whose geotag refers to a point in Philadelphia County, PA. This tweet would be classified either as local or non-local depending on whether or not the user who posted the tweet is considered to be a local of Philadelphia County. More specifically, this is how each metric would make its localness assessment:

- *n-day*: If the user tweets multiple times in Philadelphia County over a span of at least 10 days, the tweet would be considered local.
- *plurality*: If the user had posted more (or equal) tweets in Philadelphia County than any other county, the tweet would be considered local.
- *geometric median*: If the user had posted at least five tweets and enough of them were centered in the Philadelphia area for the median to be in Philadelphia County and within 30km of half of the user's points, the tweet would be considered local.

Table 3.1. Localness metric recall.

The recall of each metric, or the percentage of users who were assigned as local to at least one county.

Repo	n-days (10)	Plurality	Geo. Med.	Loc. Field
T-51M	60.10%	100.00%	49.50%	34.40%
T-11M	65.90%	100.00%	24.70%	37.30%
F-15M	67.90%	100.00%	33.90%	31.10%
S-8M	88.30%	100.00%	69.10%	15.90%

- *location field*: If the user had written in her Twitter location field "Philadelphia" or "Philly" (or a similar variant) and that text was successfully geocoded to a lat/lon in Philadelphia County, the tweet would be considered local.

3.5.2. Results: Comparing Localness Definitions

Running all four localness metrics against the same datasets affords us a unique ability to compare and contrast the definition of localness each metric encodes. Overall, three trends emerge: (1) some localness metrics fail to identify a single local county for many users, (2) though we see substantial agreement in localness determinations for the users bridging our two Twitter datasets, there is a large minority for whom results vary, and (3) although there is not strong agreement between any of the metrics, *n-days*, *plurality*, and *geometric median* agree far more often with each other than any of the three do with *location field*.

3.5.2.1. Highly-varied Recall. With regard to the first theme, efforts to filter out non-local geotagged social media have not considered recall as an issue. However, Table 3.1, which shows the percentage of users for which each metric was able to find at least one local county, suggests that this is an important factor to consider. If a localness metric is not able to find a local region (e.g., county) for a given user, that user can have no local social media, thereby removing him/her from social media VGI studies that filter for localness (a practice

strongly supported by other results in this chapter). In cases when data is not plentiful (e.g., for analyses at very granular spatial scales), limited recall could be a major problem.

Looking at Table 3.1 in more detail, we see that while *plurality*, by definition, succeeds for all users, *location field* sits at the opposite end of the spectrum, failing to identify a local country for well over half of all users in every case. *Location field*'s low recall is most likely attributable to two factors: (1) not all users fill out their location fields (especially on Swarm) and (2) location field entries are often non-geographic in nature, which will result in the geocoder returning no value (in the ideal case) [**138**].

3.5.2.2. Varying Longitudinal Consistency. Examining the set of tweets from the 389,635 users who appeared in both our *T-51M* and *T-11M* datasets, we found a very high consistency for the *location field* metric (91%) as well as *geometric median* metric (74%). In other words, for users we could identify in both datasets, the counties for which they were considered local were frequently the same using *location field* and *geometric median*, even though there is a 7-month gap between the data collection periods. However, the same is not true for *plurality* (54%) and *n-days* (48%), suggesting that there is a trade-off between recall (*plurality* and *n-days* both have high recall) and longitudinal consistency. This is a point to which we return in the discussion section.

3.5.2.3. Different Localness Definitions, Different Results. Each localness metric operationalizes a different idea of localness, and, as such, it is not a surprise that they frequently disagree as to whether an individual piece of VGI can be considered a local to a county. *Plurality, n-days* and *geometric median* agreed the most, but for instance, their output agreed that a given tweet was local only 76.9% of the time for *T-51M*, and that is the highest agreement of any of the four repositories. *Location field* rarely came to the same conclusions as any of the other three metrics, and, as such, the repository for which there

Table 3.2. Percentage of social media content that is local.

Relative percentage of social media VGI classified as local. Geotagged social media from users for whom no local county could be identified are excluded from these figures

Repo.	n-days (10)	Plurality	Geo. Med.	Loc. Field
T-51M	84.00%	90.10%	91.20%	57.90%
T-11M	77.10%	76.90%	85.00%	51.10%
F-15M	78.40%	52.90%	70.70%	40.70%
S-8M	88.20%	70.10%	73.00%	1.10%

was the most agreement across all four metrics (*F-15M*) still had only 16.3% agreement (with agreement defined the same way as above).

Because of the diversity in localness operationalization across the four metrics, the remainder of our studies below use at least two metrics, and usually use all four so as to establish robustness across varying definitions of localness. In the discussion section, we outline how this approach is likely a best practice for social media VGI research more generally.

3.6. RQ1: How Local is Social Media VGI?

In this section, we discuss our research on assessing the degree to which social media VGI is local. In other words, in asking this question, we are inquiring whether the localness assumption is valid.

3.6.1. Methods

Once we had completed our work for RQ0, addressing RQ1 was very straightforward: we simply calculated the percentage of overall social media units that are local according to each algorithm.

3.6.2. Results

Table 3.2 shows the localness of each social media VGI repository according to each of the localness metrics. The picture of social media VGI localness that emerges from Table 3.2 is that while *the majority of VGI appears to be local according to most metrics, a large minority is non-local*. For instance, we see that according to *n-days*, the localness of our four repositories ranges from 77.1% (*T-11M*) to 88.2% (*S-8M*), with the *T-51M* and *F-15M* repositories' localness between these two values. With the median localness percentage across all four datasets and all four metrics being only 75%, *it is difficult to make the argument that the localness assumption holds true in social media VGI.*

3.7. RQ2: Does Localness Vary Geographically?

In this section, we describe how we addressed our research question related to potential geographic variation in the localness results we reported above. We focus this investigation on whether any variation corresponds to important sociodemographic contours.

3.7.1. Methods

As the first step in addressing RQ2, we calculated the percent of social media VGI in each county that is local to that county. This is analogous to our approach outlined above for RQ1, but instead of identifying the share of social media that is local in entire repositories, we did so on a county-by-county basis. We then analyzed the localness ratio in each county (for each repository) in the context of key sociodemographic statistics of the county (see §3.3).

This analysis was conducted using a multivariate regression with percent local as the dependent variable and the sociodemographic statistics as independent variables. We first

Table 3.3. Localness sociodemographic regression results.

Summary of % Localness Multivariate Regressions for *n-day* (10) localness filter. *** is p<0.001, ** is p<0.01, and * is p<0.05.

Repo.	% Urban	MedAge	HMI	%WNL	%MBSA
T-51M	0.29***	-0.18***	-0.06**	0.15***	-0.05**
T-11M	0.40***	-0.14***	-0.04	0.02	-0.04*
F-15M	0.28***	-0.06**	-0.01	0.14***	0.07***
S-8M	0.39***	-0.18***	0.01	0.16***	0.02

test for spatial autocorrelation (if none is found, traditional OLS would be appropriate) and then adjust for the presence of spatial autocorrelation as discussed in [**56, 201**] by running either a spatial error or lag model from the R library package spdep [**24**]. We make the specific choice of model based on Lagrangian Multiplier measures of fit, which is considered a best practice in the field of spatial econometrics [**11**]. The dependent and independent variables are log-transformed as necessary to achieve normality and all variables are Z-score standardized so that we can relate all coefficients to changes in standard deviations and compare relative effect sizes accordingly.

3.7.2. Results

Though we see, on average, that approximately 75% of content is local, the standard deviation for most metrics and datasets is around 20%, suggesting that there is noticeable geographic variation in the degree to which content is local. The results of our spatial regressions, which describe the percentage of local content in a county as a function of its sociodemographic factors, can be seen in Table 3.3. We report only the results for *n-day* calculations, but we ran the spatial regressions for each localness algorithm and found that all four algorithms had very similar results within each repository[3].

[3]The sole exception was location field for Swarm, which had mostly insignificant coefficients for the regression due to low recall.

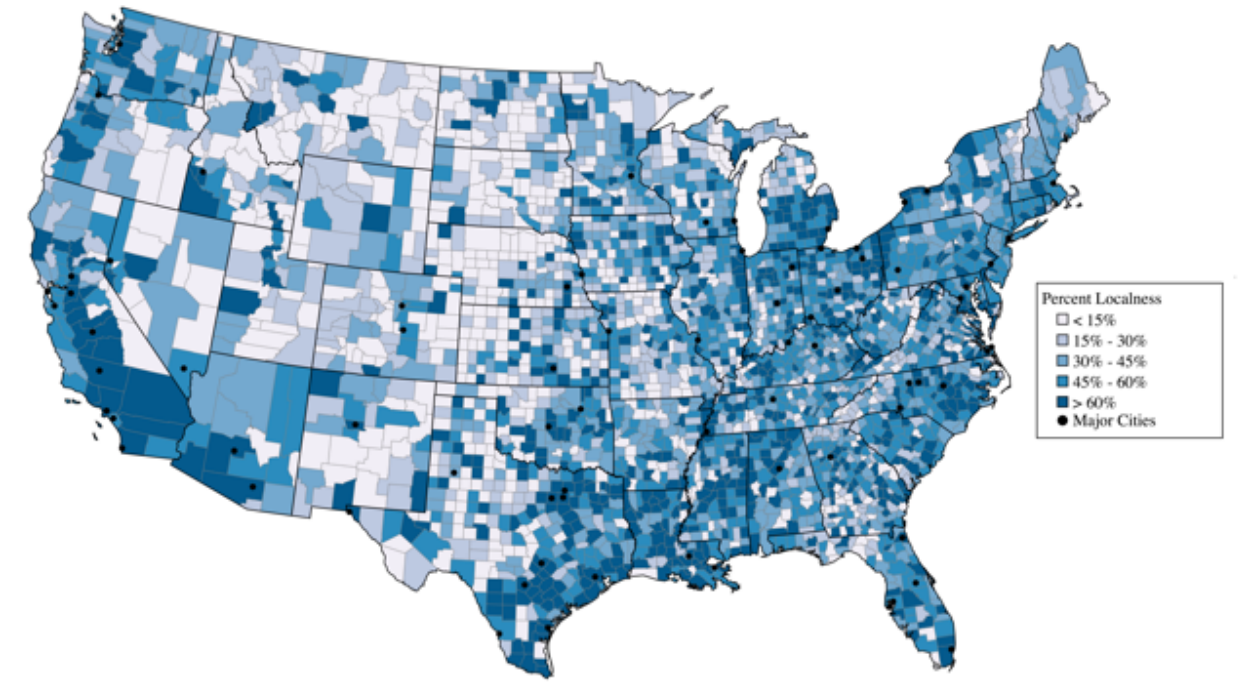


Figure 3.1. Map of Percent Local Content in T-11M according to the geometric median metric.

Table 3.3 reveals that the county-level variation in localness percentages is certainly not random; we see significant and consistent effects for a number of our independent variables. Across the board, there appear to be moderate increases in localness with increases in *%WNL* and increased youth (i.e. decreased median age) and much larger increases in localness with increased *% Urban Pop*.

Examining the effect sizes from our regression, we see that with all else held equal, for every standard deviation increase in the percent of the population that is urban (+31.6%), there is a 15-40% absolute increase in the localness of social media. This relationship results in social media VGI of substantially different character at opposite ends of the *% Urban Pop* spectrum. For instance, for counties with a *% Urban Pop* $>$ 90% (e.g., San Francisco

County, CA; Cook County, IL), 82% of *T-51M* tweets whose geotag is in the county come from a local user according to *n-day*. The corresponding value for counties with *% Urban Pop* $< 10\%$ (e.g., Twin Falls County, ID) is only 63%. Figure 3.1 shows a similar trend occurring with geometric median across the *T-11M* dataset, with the very rural Great Plains region containing much lower proportions of local content than more heavily populated areas. A similar trend occurs with *Median Age*, but with a smaller magnitude. Unpacking the normalized effect sizes in Table 3.3, there is a 6% decrease in localness in *T-51M* as the median age shifts from 32 (e.g., Newport News, VA) to 47 (e.g., Hernando County, FL).

A striking trend in Table 3.3 is the extent to which the table largely mirrors known findings about overall population bias in our three social media communities. It appears that not only is there more social media VGI per capita in urban areas [**140, 201**], but our results suggest that urban social media is also far more local. The same is true with regard to population biases in age; older areas have less social media VGI per capita, and it is less local. This result is a point to which return in the discussion section below.

The results in Table 3.3 also suggest that the importance of filtering for localness is not geographically uniform. It appears that adopting the localness assumption will reduce the accuracy of studies that use social media VGI in rural areas more than it will reduce the accuracy in urban areas. The same is true of older areas vs. younger areas, areas with a smaller WNL population vs. a larger one (see §3.9), and so on.

3.8. RQ3: Impact of Non-Local VGI

While RQ1 and RQ2 sought to characterize localness in social media VGI, RQ3 seeks to directly understand its effects on social media VGI-based research. To do so, we adopt a case study approach, focusing on an analysis performed by Mitchell et al. [**223**]. This analysis

used their algorithm from [**64**] applied to geotagged tweets to calculate the geography of happiness in the United States.

3.8.1. Methods

Using the data and code for their algorithm provided by Dodds et al.[4] and a combined version of our two Twitter datasets (*T-51M + T-11M*), we computed the geography of happiness in the United States at both the state-level and the county-level. We performed this computation three times: once under the localness assumption (in which we did no filtering for non-local tweets), once filtering out non-local tweets using *n-days* and once doing the same with *plurality*[5]. By comparing the results of these three computations, we can gain an understanding of the effects of the localness assumption (as well as effects of filtering by each localness metric).

We include the 48 contiguous states as well as Washington D.C. in our states calculation but limit the counties results to only those counties containing at least 3000 total tweets to ensure a sufficient sample size of tweets for the happiness algorithm.

3.8.2. Results

At the scale of states (e.g., *n-days* and *plurality* would group contributions from Los Angeles together with those from Yosemite National Park), we see no major shifts in the rankings of happiest states when filtering on *n-days* or *plurality* (i.e. when filtering out non-local tweets according to those metrics). The top three happiest and top three saddest states for each

[4]https://github.com/andyreagan/labMT-simple

[5]Following the approach implemented by Hecht and Stephens, instead of removing non-local tweets as we have in RQ0 through RQ2, we instead assign them to their local counties/states.

Table 3.4. States ranked by happiness.

The happiest and saddest states using an unfiltered dataset (i.e. no localness metric applied), *n-day*, and *plurality*.

Ranking	Unfiltered	n-day	Plurality
1	Montana	Montana	Montana
2	Vermont	Vermont	Maine
3	Maine	Maine	Vermont
...		...	
47	Delaware	Mississippi	Mississippi
48	Maryland	Maryland	Louisiana
49	Louisiana	Louisiana	Maryland

Table 3.5. Counties most affected by localness.

Counties with some of the largest shifts in happiness ranking after filtering out non-local tweets.

County	Unfiltered	n-day	Change
St. Louis County, Missouri	162	657	-495
Baltimore County, Maryland	983	1251	-268
San Francisco, California	139	309	-170
Mercer County, West Virginia	475	296	179
Iroquois, Illinois	479	289	190

implementation are shown in Table 3.4. The one notable shift that we see is that tourism-heavy Nevada moves from a relatively happy rank 15 out of 49 in the unfiltered dataset all the way down to middle-of-the-road rank 25 when non-local tweets are filtered out, by far the largest shift of any state.

At the scale of counties, the difference in ranks between the unfiltered computation and those using *n-days* and *plurality* is larger, but not tremendously so. For *n-days*, the median absolute change in ranking from the unfiltered rankings is 32, but there are 105 counties that see shifts of more than 10% (126 rankings) up or down the list. *Plurality* had very similar results (median change 32; 96 moved by 10%).

The high-level results, however, obscure an interesting phenomenon in which several counties experienced very large jumps up and down the ranks when localness filtering was

introduced. Examples of some of these counties are shown in Table 3.5. These are counties where the signal coming from the local populace is being overwhelmed by a very different signal from the non-local population. The greatest single change occurs in St. Louis County, Missouri, with a drop of 495 rankings when non-locals were filtered out. Baltimore County, Maryland, also moved substantially (10^{th} most) with a drop of 268 rankings. During the data collection periods, these two counties were experiencing fallout from the deaths of Michael Brown and Freddie Gray, respectively. In both cases, it appears that while local sentiment declined precipitously, this decline was obscured in the unfiltered dataset by travelers (non-locals), a group that did not experience the drop in sentiment.

The reverse appears to be happening in Mercer County, Arkansas, and Iroquois County, IL, which are both rural areas that happen to lie on major interstate highways. In both instances, non-locals who were driving through likely caused the drop in happiness in the unfiltered dataset relative to the filtered one.

3.9. Discussion and Implications

3.9.1. Best Practices for Social Media VGI Research

Ruths and Pfeffer [**265**] recently called for "higher methodological standards" for "large-scale studies of human behavior in social media." In doing so, they laid out a framework of best practices for working with social media, e.g., accounting for population biases, filtering for nonhuman accounts, and showing results from multiple platforms or temporally-separated datasets[6].

[6]We implement most of these suggestions in this work (i.e. removing non-human accounts in Twitter, testing on multiple repositories and datasets, and not relying on a single algorithm).

The research presented above points to an extension of Ruths and Pfeffer's framework that is specific to geotagged social media. Specifically, our results suggest the following five best practices:

Best Practice #1: As we found that a large minority of geotagged social media is not local, *studies that utilize geotagged social media should not adopt the localness assumption.* Doing so will result in significant and sociodemographically-biased side effects that, as we saw in our work related to RQ3, can alter study results.

Best Practice #2: We revealed clear differences between the localness metrics that have been used in the literature, indicating that *researchers need to think carefully about how to best operationalize localness for their research questions.* The decision of how to operationalize localness should involve both semantic and practical considerations, considerations that we unpack below.

Best Practice #3: With regard to semantic considerations, researchers should carefully examine which localness metric best fits the needs of their study. Each metric we identified in the literature defines localness differently, with metrics like *n-days* assuming a user can be local to many counties and metrics like geometric median assigning users to a single lat/lon coordinate. These definitional differences were likely a major cause of the deviations between the outputs of each localness metric, indicating that choosing an incorrect definition of localness may be costly.

Best Practice #4: With regard to practical considerations, according to the needs of a given study, *researchers must negotiate the trade-off of the consistency of metrics such as geometric median with the recall of metrics like plurality.* While geotagged social media occurs at massive scales at a high-level, if one is trying to study phenomena that occur in rural areas or at very granular spatial scales (or using a repository that does not make its

data as available as Twitter, e.g., Swarm), a 56% (geometric median) or 70% (location field) reduction in the amount of data that can be considered (let alone that can be assured to be local) is highly problematic.

Best Practice #5*: Wherever possible and appropriate, researchers should consider multiple definitions of localness in parallel*, as we have done here. Given the differences in what is considered local and non-local by each metric, using multiple metrics can ensure that findings are robust against different definitions of localness. To ease this burden, future work could explore whether a single ensemble metric, such as a combination of *n-days* and *plurality*, is more robust to dataset variation.

While there are geotagged social media best practices outside the localness domain that also likely need to be encoded (e.g., guidelines for geocoding location field information as in [**138**], our results make clear that a smart and intentional approach to handling localness is an important step towards a robust social media VGI study.

To promote these best practices and provide full access to our study, we have released our implementation of the localness metrics, R code for spatial regressions, additional maps, and full results[7].

3.9.2. Localness Compounds Population Bias

Population bias is a known concern with social media research. As we have shown, localness also tends to be lower in these areas already known to be disadvantaged in social media representation (e.g., older and more rural), compounding the existing population biases. In particular, this indicates that there are further barriers to robust research on rural areas or other underrepresented populations using geotagged social media. There is an established

[7]https://github.com/joh12041/chi-2016-localness

thread of geolocation inference research that seeks to locate social media without explicit location information. Our work motivates the need to develop these tools specifically with the goal of translating the non-geotagged social media of these underrepresented populations into VGI.

3.10. Limitations and Future Work

3.10.1. Scale- and Time-Dependence

While this is the first research to survey the localness metrics in use within the literature and directly compare their results, similar studies should be conducted varying the spatial scale at which localness is defined instead of the metric. It has long been known in geography that processes that operate one way at one scale may not operate in the same way at a different scale. It is possible that localness has a scale-dependent component. We began to explore this question with our *Happiness* case study, where we saw that the inclusion of non-local content at the state scale impacted the rankings to a lesser degree than at the county scale. Along with spatial scale, future work should look at the relationship between time and localness. Localness is not a time-invariant measure – that is, the degree to which one is local to an area degrades not only with distance, but also with time.

3.10.2. New Demographic Contours Should be Considered

While doing early work on the relationship between gender and localness, we noticed an interesting new difference between our localness metrics: Leveraging well-known gender inference approaches [**56, 222**], we saw a consistent-but-small female skew in the users for whom *n-days*, *plurality*, and *geometric median* were able to assign at least one local county. However, this skew flipped and strengthened significantly for *location field*, to the point that

location field is 23% more likely to be able to assign a local county to a male user than a female user. This result has potentially important implications for localness research, as well as for use of the location field on Twitter more generally (e.g., Does the use of this field induce a population bias?), implications that need to be explored in future work.

3.10.3. Establishing a Ground Truth

Finally, although we explored in detail the four major families of localness metrics in the literature, one metric that has yet to be considered but could lead to important innovation in this area is that which involves a traditional ground truth. By asking social media users where they believe they are local (outside the social context of the location field and without its particular constraints as outlined by Hecht et al. [**138**], it should be possible to construct a learned model (that uses the four localness metrics as features) of more robust self-reported localness. Moreover, adopting ground truth approaches could enable models that operationalize highly diverse understandings of localness, including those that take time into account (in a weighted or an unweighted fashion), implement a fuzzy definition of localness, and so on.

3.11. Conclusion

In this chapter, we performed the first focused exploration of the extent to which geotagged social media can be considered to be from a local to the region of its geotag, an assumption that underlies many studies that utilize geotagged social media. We find that this assumption does not hold for about 25% of geotagged social media, although the exact percentage varies from social media community to social media community. We also saw that the degree of localness varies extensively geographically, and it does so in a fashion that

mirrors existing sociodemographic contours (e.g., more rural areas are less local). Through a case study, we demonstrated that including non-local social media in research studies can lead to incorrect conclusions in certain cases, and we outlined a series of best practices to help researchers avoid this outcome and further strengthen their work.

CHAPTER 4

Peer Production and the Urban-Rural Divide

In this chapter, we consider Wikipedia and OpenStreetMap content in urban and rural areas. We compare the relative quantity of this content across these regions, how it was produced, and the resultant quality.[1]

4.1. Introduction

Peer-produced content has been a game-changer in the tremendously important domain of geographic information. We now learn about places near and far by reading peer-produced Wikipedia articles, and many of the maps we use on a daily basis leverage peer-produced data from OpenStreetMap (OSM), the "Wikipedia of maps" [**90, 91, 150, 319**]. Behind the scenes, many important intelligent algorithms utilize data from Wikipedia and OSM to make geographic inferences about the world (e.g., [**114, 215**]).

The importance of geographically-referenced peer-produced content, also known as peer production volunteered geographic information or *peer production VGI* [**139, 274**], has led some researchers to inquire as to whether the peer production content generation model works equally well in describing all types of geographies (e.g., [**113, 135, 207, 250**]). However, missing from this literature is a robust analysis of the effectiveness and character of peer production across the urban-to-rural spectrum. Researchers in HCI have shown that urban/rural dynamics can play prominent roles in a variety of online social systems ranging

[1]The work presented in this chapter was originally published in: **Johnson, I.**, Lin, Y., Li, T., Hall, A., Halfaker, A., Schning, J., and Hecht, B. Not at Home on the Range: Peer Production and the Urban/Rural Divide. *ACM Conference on Human Factors in Computing Systems 2016.*

from social networks [**105**] to photo-sharing sites [**140**] to check-in platforms [**140**]. These results echo decades of research in the social sciences that chronicles differences in how urban and rural areas have adopted and used technology (e.g., [**49, 173**]). The previous chapter (3) further demonstrated that not only is their less participation on social media in rural areas, but also a lower proportion of the content that does exist is "local". This was shown to be particularly important in the context of research, where filtering explicitly for localness changes the conclusions that one might reach about an area based on its associated social media content.

The goal of this chapter is to examine peer production's relative ability to represent urban and rural areas. We focus not just on quantity of content, but also on who generates this content and its quality to users. Because our goal is to understand urban/rural dynamics in peer production generally rather than in a single type of peer production community, we consider both Wikipedia and OpenStreetMap. These are two of the largest and most impactful peer production communities and are communities with substantially different approaches to peer production. For the same reason, we also examine content about countries with two very different human geographies: the United States and China, which are often the focus of cross-cultural analyses (e.g., [**330**]). Studying urban/rural content generation in diverse online and offline communities allows us to gain a richer understanding of the phenomena of interest [**12**].

We find that *regardless of the peer production community or country, content about rural areas is of substantially lower quality than urban areas.* For instance, Wikipedia articles about rural places are much more likely to be assessed as low quality by Wikipedia contributors than articles about urban places, rural OpenStreetMap entities (e.g., buildings, roads, etc.) have fewer tags than urban entities, and Wikipedia articles about rural areas have substantially

more content written by editors who have not specialized in the local area. Indeed, our results show that Wikipedia articles about very rural places in the U.S. have less than 5% of their content (on average) written by specialized editors, whereas this number is over 37% in urban areas.

Our results also highlight the important role of "bots" (i.e. automated software agents) and automation-assisted batch editors in creating content about rural areas. For instance, whereas 4.5% (median) of content in Wikipedia articles about urban places is bot- or batch editor-generated, the corresponding proportion for rural places is over 23%. Although these automated solutions generate low-quality content, we show how they are critical to peer production's ability to provide a baseline level of coverage in rural areas.

Rural areas have significantly different sociodemographic characteristics than urban areas, and the models we use in this research control for these potential confounds. Through these modeling efforts, this research also tangentially uncovers new findings about systemic peer production biases in our control variables. For instance, we find that English Wikipedia articles about places with a higher percentage of Democratic votes are of higher quality, and the same is true of more educated places in the Chinese Wikipedia.

As we will detail below, these results have highly tangible implications for (1) human consumers of peer-produced content (e.g., Wikipedia readers, users of OSM-based maps) and (2) the many systems and algorithms that rely on peer-produced content to understand the world's geography. With regard to humans, our results mean that articles about rural areas and OSM-based maps about rural areas are of substantially lower quality (on average) than those about urban places. With regard to systems and algorithms, our results suggest that, to many peer production-based geographic technologies, rural areas frequently "look the same" and are defined only by their topology and government census data.

This research is also the first to characterize and quantify the fundamental challenge facing peer production communities' efforts to map and describe rural phenomena: in rural areas, the ratio of entities of interest (e.g., incorporated towns) to potential local editors can be orders of magnitude higher than in urban areas. Since many contributors of geographic peer-produced information contribute about places local to them (e.g., [**128, 136**]) and local contributions are generally of higher quality (e.g., [**73**], our coding study below), rural areas are systemically disadvantaged in the peer production content generation model. We show how this disadvantage manifests itself in the results we identify in this chapter. Fortunately, our results also point to potential partial solutions to this issue, and we expand on these implications for design below.

To summarize, this chapter makes the following contributions:

- We establish that peer-produced content about urban areas is of higher quality than that about rural areas, demonstrating that this difference persists across two prominent peer production communities (Wikipedia and OpenStreetMap) with very different communication and collaboration structures and two countries with very different human geographies.
- We highlight the critical role that bots and batch editing tools play in ensuring that there is any content at all about some rural areas.
- We discuss how our results reveal systemic challenges facing peer production projects in describing and mapping rural phenomena and highlight how our results can inform the design of potential sociotechnical solutions.
- Finally, through controlling for sociodemographic factors, we identify new content biases in peer-production related to politics, education, and profession.

4.2. Related Work

This chapter draws motivation from work in three areas: (1) coverage biases in peer production, (2) urban/rural dynamics in online social systems, and (3) content localness in peer production. Below, we discuss each of these areas in more detail, with other related work discussed in context in later sections.

4.2.1. Coverage Biases in Peer Production

This research is informed by a thread of related work that looks at systemic variations in peer-produced content along dimensions other than the urban/rural spectrum (see §2.3.1 for an overview). A major recent thrust of this work relates to gender dynamics, with several studies showing that both OSM and the English Wikipedia have more content about and for men than about and for women (e.g., [**140, 181, 217, 312, 256**]). Language has been shown to be a particularly strong factor in Wikipedia coverage biases, with each language edition having much better coverage of places where the corresponding language is spoken [**135, 274**]. Other factors behind systemic variations in content include national culture [**250, 249, 241**], and politics [**115**].

4.2.2. Urban/Rural Dynamics in Online Social Systems

The vast majority of work that considers issues associated with online social systems and population density has focused exclusively on urban areas [**105, 140**]. The small body of online social systems research that has looked at differences between urban and rural areas has largely focused on social media, which has significantly different collaboration and contribution characteristics than peer production. For instance, Gilbert et al. [**105**] found

that rural MySpace users had significantly fewer friends than urban users. Similarly, Hecht and Stephens [**140**] identified that rural areas have fewer tweets, check-ins, and geotagged Flickr photos per capita than urban areas.

In work that provided key motivation for the research in this chapter, several researchers have observed that OpenStreetMap data seems to display different characteristics in rural and urban areas. For instance, Zielstra et al. [**356**] identified that OSM coverage was more extensive closer to a selection of German cities, and a similar finding was identified in the London area by Mashhadi et al. [**207**]. One goal of this chapter is to build on these findings, most importantly by examining OSM and Wikipedia in a consistent, robust analytical framework so as to gain an understanding of peer production in rural areas as a whole, but also by focusing specifically on urban/rural dynamics and doing so across entire countries. This allows us to incorporate important control variables, use more sophisticated urban/rural sociodemographic statistics, and introduce other targeted approaches and metrics, ultimately resulting in the series of contributions with important implications outlined above.

4.2.3. Localness and Peer Production

Important context for this work also comes from the small literature on the localness of peer-produced content. This literature has shown that, in general, edits to geographic Wikipedia articles tend to come from people who are located (and likely live) near the subject of the article. For instance, Hecht and Gergle [**136**] found that over 25% of edits to Wikipedia articles about places in the English Wikipedia comes from within 100km and a similar finding was identified by Hardy et al. [**128**]. Related research on OpenStreetMap has shown that local editors tend to do higher quality work (something we confirm in our studies below). For example, Zielstra et al. [**354**] found that OSM editors contributed a higher diversity of

edits in their home region and Eckle [**73**] found that familiarity with an area led to more accurate OSM mapping in a controlled experiment.

Recent work by Sen et al. [**274**] examined the geographic variability in these overall localness results, finding that Wikipedia articles about places in poorer countries and countries with a less healthy publishing industry have fewer local editors and reference fewer local sources. A portion of our research extends the work of Sen et al. to a more granular scale, looking at whether localness also varies across the urban/rural spectrum and finding that rural places have less local content, even on a per capita basis.

4.3. Data and Metrics

This research involves a number of different datasets and a variety of metrics, with many of these datasets and metrics being the output of somewhat complex processes. We first describe basic information about our Wikipedia and OSM datasets. Next, we discuss each of our other datasets and metrics in detail, grouped by whether they help us quantify (1) rural/urban dynamics, (2) peer-production quality, (3) peer-production quantity, or (4) control sociodemographic variables.

4.3.1. Peer Production Datasets

Most of our Wikipedia data was extracted from the static XML dumps of the English and Chinese language editions of Wikipedia using the WikiBrain API [**273**]. We analyzed the English Wikipedia when considering the United States and the Chinese Wikipedia when considering China. Like nearly all prior geographic Wikipedia work (e.g., [**113, 135, 136, 193**]), we focus on "geotagged articles", or articles that have been tagged with a latitude and longitude location by Wikipedia editors. In total, we identified 218,709 English geotagged

articles about places in the contiguous United States and 46,124 Chinese geotagged articles about places in China. Note that because of the requirements of our geographic modeling techniques (see next section), we only consider the contiguous 48 states when examining the United States.

Our OpenStreetMap data comes from the static Planet.osm dump from February 2014 for the United States and July 2015 for China. For the contiguous US, the dump contains 494 million nodes, 32.8 million ways, and 263 million tags. In China, there are 22.6 million nodes, 1.6 million ways, and 5.5 million tags.

We aggregate all Wikipedia and OpenStreetMap metrics to the county level in the contiguous United States (3,109 counties) and the prefecture level in China (344 prefectures). The census data necessary to perform our analyses at a more local scale in China is not available from the Chinese government (see below). We filter out geotagged Wikipedia articles about entities with footprints larger than a county/prefecture (e.g., articles about states, countries, continents) to avoid assigning their content to the single county/prefecture that contains the lat/lon of their geotag. It is important to point out that this aggregation allows us to make claims about *all data about places in a country/prefecture*, rather than the specific Wikipedia article about a county/prefecture or the specific OSM geometry describing the county/prefecture. This aggregative approach has often been shown to be effective in related work (e.g., [**113, 135, 207, 249**]).

4.3.2. Urban/Rural Datasets

We obtained data about geographic variation in urbanness from government sources. In the United States, we make use of a statistic from the 2010 U.S. Census that describes the percentage of the population in a given county that lives in an urban area (*% Pop*

Urban). This percentage is calculated using the U.S. Census' definition of urban areas, which includes both significant cities as well as urbanizations of 2,500 or more people [**35**]. When our analyses require discrete classifications along the urban/rural spectrum, we utilize the National Center for Health Statistics' (NCHS) urban-rural classifications [**151**], which assign each U.S. county an ordinal code from "1" (core urban) to "6" (entirely rural). New York County is assigned a "1", for example, and Loving County, Texas (the lowest-population county in the United States) is assigned a "6". In China, our *% Pop Urban* statistic comes from the Chinese government's 2010 Population Census. Urban areas are defined based on administrative districts and mainly include highly commercialized and populous districts.

4.3.3. Content Quality Data and Metrics

There are many definitions of content quality in peer-produced datasets, with each definition serving an important constituency and providing a unique view into the effectiveness of the content. In order to gain as deep an understanding as possible of quality variation across the urban/rural divide, we sought to examine both repositories using a variety of quality definitions aimed at uncovering different dimensions of peer production quality. Many quality metrics are country- or repository-specific due to the requirements of the methods by which they are calculated.

4.3.3.1. WikiProject Quality Assessments. In the English Wikipedia, most articles are assessed by members of the Wikipedia community with a quality score from an ordinal seven-point scale that ranges from "stub" class ("provides very little meaningful content") to "featured article" class ("a definitive source for encyclopedic information") [**322**]. These assessments have been used in a number of research projects as a holistic measure of the multifaceted notion that is Wikipedia article quality (e.g., [**317, 316, 318, 169**]). We

evaluate the quality differences between urban and rural areas using this rich quality metric by measuring the percentage of articles about places in a U.S. county that are assessed at C-class or higher (*% C-class or higher*). We use C-class as the threshold as Wikipedia's documentation describes it as the lowest quality class in which articles are still "useful to a casual reader" [**322**]. While the Chinese Wikipedia does have an analogous quality scale, it has not been validated in the literature to our knowledge, and, as such, we restrict this metric to the English Wikipedia/United States.

4.3.3.2. Tag Richness. In addition to describing the geometries of geospatial entities, OpenStreetMap also contains a large dataset of tags corresponding to these entities. OpenStreetMap tags are how humans and computers understand the semantics of the underlying geometries. Without them, maps (and algorithms) based on OSM data would not be able to, for instance, distinguish a hospital from a bar or a highway from a dirt road [**293**]. Tags also support OSM-based location-based services by providing them with venue opening hours information, cuisine type, and many other attributes. In general, the more tags an entity has, the more useful it is to humans and computers. To operationalize this notion of quality, we use a metric called "*Tag Richness*", which is defined simply as the average number of tags per entity in a county/prefecture.

4.3.3.3. Content Diversity. One quality metric that plays an important role in the value both humans and computers gain from peer-produced content is the amount of unique information available about a specific place. For instance, boilerplate Wikipedia articles about a town that merely describe the town's neighboring towns and basic census statistics are less useful to readers and algorithms/systems than articles that have rich descriptions of the town's unique history and character.

While the value of diverse content to readers is obvious, the value for systems/algorithms is more complex (but equally important). Systems/algorithms that use peer-produced knowledge typically use data models derived from article/region content (e.g., "bag of links" models for Wikipedia [**221**]), leveraging these data models to answer queries, assess the similarity of concepts, and support many other applications (e.g., [**323, 93**]). If articles all have roughly the same content, the power of these systems/algorithms to discriminate between different places will be adversely affected, likely reducing application effectiveness in rural areas in some cases.

To operationalize content diversity, we use a metric we call *Outlink Entropy*, which has the advantage of being directly linked to the diversity of commonly-used "bag of links" models as well as capturing a human-visible notion of content diversity. A straightforward application of information entropy, outlink entropy measures the extent to which the links on pages about places in a county/prefecture all point to the same small set of articles, or whether they point to a diverse group of articles. For instance, if a large proportion of links in a county's geotagged articles point to the "United States Census" article (because a large proportion of the articles' content amounts to basic census statistics), this would result in low outlink entropy. In a higher entropy county, articles' links would point to a diverse set of other articles relevant to the county's history, current events, and so on.

4.3.3.4. Ratio of Human-generated Content. As has been described in prior work by Geiger and others (e.g., [**101, 102, 122**]), peer production communities are often complex ecosystems that consist of human editors and automated and partially-automated software agents. These agents "promote consistency in the content, structure, and presentation" of articles [**295**], and, in some cases, generate content. Much of this automated content generation usually amounts to the importing of pre-existing data or statistics into the genre

and format of the peer production community. For example, many geotagged Wikipedia articles have text that a bot generated from census statistics, e.g., in the "Clayton, Missouri" article, there is bot-produced text that reads "As of the census of 2010, there were 15,939 people, 5,322 households, and 2,921 families residing in the city." Similarly, in OSM, large quantities data have been imported from the U.S. government's TIGER/Line street dataset.

The content generated by automated and partially-automated agents is often considered by members of peer production communities to be of substantially lower quality than that generated by humans. Indeed, in both OSM and Wikipedia, extensive debates have taken place as to whether automated content generation should continue, and, if so, whether and how it should be constrained [**194, 355**]. To capture this notion of quality, we measured the amount of content contributed in each county/prefecture by human editors versus that contributed by automated or partially-automated agents.

Following prior work on identifying bots in Wikipedia [**316**], we distinguished human Wikipedia editors from bot editors by comparing the editor's username to usernames in the "bot" user group as well as by searching the username for the word "bot". Batch editors are identified by doing a case-insensitive search for the names of two very common editors, "AWB" and "WPCleaner", an approach that has been used successfully in prior work [**167**].

In OpenStreetMap, a feature was identified as bulk uploaded if its most recent edit was from a changeset in which edits occurred at a rate of faster than one per second and at a volume greater than 1000 edits. This approach is similar to prior work [**355, 354**]. We report values from this classification but also tested our data with more relaxed criteria, which produced very similar results. Notably, using this metric, features that were initially uploaded in bulk but have since been edited in a sufficiently small or slow changeset are classified as human-edited.

Measuring the amount of content attributable to a specific editor or class of editors (e.g., bots vs. humans) in Wikipedia is non-trivial. Most past work has used the number or ratio of edits, but edits can be of different sizes, can be of malicious intent (e.g., vandalism), and so on. This is a known issue in the literature, and, to address it, we turn to the work of Halfaker et al. [**121**] in tracking the persistence of words through revisions of Wikipedia pages. We process the entire edit history for each geotagged article and compute the percentage of tokens (i.e. words, numbers) in the final version of the page that were contributed by each type of contributor (bots, batch editors, and humans).

4.3.3.5. Content Contributed by "Local Experts". An important recent thread of Wikipedia research has adopted as a quality metric the extent to which the content in geotagged articles is coming from local experts (e.g., [**113, 274**]). This research is motivated by recent studies (e.g., [**73, 354**]) and by prominent geographer Michael Goodchild's claim in his formative article on VGI that "the most important value of volunteered geographic information may lie in what it can tell us about local activities" [**111**].

All prior localness work in the Wikipedia domain has relied heavily on IP geolocation, and nearly all of this work has examined localness at the country-to-country scale (i.e. a contributor is "local" if she is from the same country as the entity she is editing/contributing). Because the accuracy of IP geolocation declines tremendously when attempting to position IP addresses to their state, county, city, etc. rather than their country [**244**], the approaches for quantifying local expertise in prior work cannot be used for our more granular analyses (analyses that have been called for by some of this prior work, e.g., [**274**]).

To address this problem, we instead assess the percentage of Wikipedia article tokens about a given U.S. county that come from contributors who have exhibited some degree of local focus on that county. This has the benefit of allowing an editor to be considered a

"local expert" in more than one county (e.g., their home county and the county in which they attend university), while at the same time filtering out edits by "fly-by" editors (we compare "local focus" and "fly-by" edits below). Specifically, as our definition of local expertise, we measure the percentage of tokens about a county that come from editors who have focused 10% or more of their effort on that county (as measured by edits to geotagged articles). This more multifaceted definition of localness is similar to the "n-days" metric that has been used when studying urban/rural dynamics in social media VGI [**140, 187**]. However, because this definition is not directly comparable with past definitions of local expertise, we do not describe this quality metric as a "local expertise" metric, but rather a metric that measures the degree of "local spatial focus".

To confirm that editors who display local focus contribute different types of information than fly-by editors and to better understand the nature of each group's contributions more generally, we performed a small qualitative coding exercise. Two coders examined all tokens contributed by editors with local spatial focus and fly-by editors on 25 randomly selected articles about places in NCHS = "6" counties. The coders classified these tokens into four categories: (1) bot-like structured data (i.e. tokens that describe data from large, well-known external repositories like the U.S. Census), (2) structured data from local sources (i.e. tokens that describe data from a very local government agency), (3) administrative edits (e.g., typo fixes, syntax fixes), and (4) rich local information (e.g., information about the area's history, culture, or well-known persons).

The coders overlapped on tokens for five articles and achieved a relatively good Cohen's $\kappa = 0.69$. As expected, the proportion of tokens that contained rich local information was much higher for editors with a local focus (61.3%) than for fly-by editors (24.0%) ($z = 5.65$,

$p < 0.001$). Moreover, over 39% of fly-by editors' tokens were bot-like in nature compared to just 16% in those with a local focus ($z = 3.39$, $p < 0.001$).

4.3.4. Content Quantity

In addition to assessing variations in quality across the rural/urban spectrum, we also examine variations in the raw amount of content across this spectrum. We do so for several reasons. First and foremost, by examining the number of entities in urban and rural areas, we can assess the per capita "burden" in rural areas relative to that in urban areas. Secondly, the vast majority of existing work on biases in peer production repositories – especially those in the Wikipedia domain – has looked exclusively at quantity metrics (e.g., number of edits per capita in sub-Saharan African countries vs. Europe [**113**], length of articles about women vs. length of articles about men [**256**]. As such, analyzing the variation of these metrics in rural and urban areas has two additional benefits: (1) it affords comparability with this existing work and (2) as we will see, our results will reveal flaws in using raw content quantity alone as a metric in peer production bias research.

We selected our specific content quantity metrics by identifying metrics that are commonly used in the peer production bias literature (e.g., [**113, 181**]). In Wikipedia, we look at number of articles per capita, number of outlinks per capita, and article bytes (length) per capita. In OSM, we examine nodes (points) per capita, ways (lines and polygons) per capita, and total tags per capita.

4.3.5. Sociodemographic Control Variables

In both the U.S. and China, the human geography of rural and urban areas has sociodemographic differences that go well beyond population density. For instance, in the U.S.,

rural areas tend to be older, poorer, and vote more Republican [**173**]. In China, rural areas are poorer, less educated, and have a higher proportion of males [**283**]. In this research, we adopt two parallel perspectives on these associations. The first attempts to control for these factors, teasing out a purer effect for rural and urban (using *multivariate* models). The second considers rural areas as they are today (e.g., on average poorer, older, more Republican), incorporating their entire human geography in our assessment of variation in peer production content across the urban and rural spectrum (using *univariate* models).

The specific sociodemographic controls we consider in the U.S. are household median income (*HMI*), median age (*Median Age*), the percent of the population that is White and non-Latino (*% WNL*, a commonly used statistic in race and ethnicity work), the 2012 vote rate for Obama (*% Democratic*), and the percent of the population employed in management, business, science, or the arts (*% White Collar*). These data come from the 2010 U.S. Census (*% WNL*, *Median Age*), the 2009-2013 U.S. Census American Community Survey (*HMI*, *% White Collar*), and The Guardian (*% Democratic*). The controls we included for China are gender ratio (*% Male*), the percent of the population that is not of Han ethnicity (*% Non-Han*), the percent of the population that is 15-64 years old (*% Age 15-64*), and the percent of the population that is college-educated (*% College or More*). All these data come from the 2010 Sixth National Population Census.

Other potential confounds (e.g., education in the U.S.) were not possible to include because of excessive collinearity with existing variables that would have destabilized the model coefficients or caused excessive positive skew, as with the very high broadband penetration rates in the United States. We also note that we included a dummy variable reflecting the presence of land managed by the National Park Service initially as a control (e.g., national parks). We anticipated that this information would help to distinguish between rural areas

of two significantly different functions and characters. However, we found including this control only minimally changed the effects of the other variables and therefore removed it from the analysis framework.

4.4. Methods

Once the metrics described above had been calculated or collected, the remainder of our methodological approach consisted of a relatively straightforward univariate and multivariate regression-based modeling exercise (with the exception that our models need to account for spatial autocorrelation; see below). Our peer production quality and quantity metrics are our dependent variables and *% Pop Urban* and the other sociodemographic variables are our independent variables. We ran a separate regression for each dependent variable.

We log-transform variables as necessary to achieve normality. We then z-score scale all variables so that the resulting beta coefficients as produced by the regressions are directly related to unit standard deviation changes in the dependent variable. This approach allows for comparison of relative effect sizes between different variables. We test for spatial autocorrelation and run spatial regressions using the *spdep* package in R [**24**] according to spatial statistics best practices, which call for selecting one of two spatial regression models (error or lag) with fit test statistics [**11**].

4.5. Results

Table 4.1 contain the results of our spatial autoregressive models. Each row in Table 4.1 corresponds to one of the quality and quantity metrics defined above, and the cells of the table are populated with normalized effect sizes for the independent variables.

Table 4.1. OSM and Wikipedia regression results.

The results of our univariate and multivariate regressions for Wikipedia and OSM in both the United States and China. Each cell value represents the corresponding normalized beta coefficient. *** is $p < 0.001$, ** is $p < 0.01$, and * is $p < 0.05$. Abbreviations: "WNL" is "White, Non-Latino"; "W.C." is "White Color"; "pc" is "per capita"; "QL" is Quality; "QN" is Quantity.

United States

Dataset / Metric			Univariate	Multivariate					
Platform	Attribute	Type	%Urban	%Urban	HMI	MedAge	%Dem	%WNL	%W.C.
Wikipedia	Outlink Entropy	QL	0.34***	0.29***	-0.07***	-0.01	-0.10***	-0.12***	0.10***
Wikipedia	% ≥ C-Class Articles	QL	0.31***	0.27***	0.00	0.00	0.15***	0.03	0.13***
Wikipedia	≥ C-Class Articles pc	QL	0.13***	0.12***	-0.03	0.03	0.20***	0.10**	0.09***
Wikipedia	% Local Focus Tokens	QL	0.29***	0.24***	-0.04*	-0.02	0.10***	0.06***	0.15***
Wikipedia	Local Focus Tokens pc	QL	0.11***	0.09***	0.03	0.04*	0.13***	0.09***	0.06***
Wikipedia	% Human Tokens	QL	0.31***	0.26***	-0.02	-0.01	0.08**	-0.03	0.12***
Wikipedia	Human Tokens pc	QL	-0.42***	-0.37***	-0.17***	0.17***	0.20***	0.08**	0.15***
OSM	Tags per Feature	QL	0.20***	0.18***	-0.07***	-0.05**	-0.01	0.05**	-0.01
OSM	% Human Nodes	QL	0.32***	0.26***	0.10***	-0.05*	0.15***	0.12***	0.13***
OSM	% Human Ways	QL	0.30***	0.21***	0.09***	-0.09***	0.17***	0.10**	0.05**
Wikipedia	Articles pc	QN	-0.51***	-0.42***	-0.16***	0.19***	0.14***	0.06*	0.07***
Wikipedia	Length (Bytes) pc	QN	-0.47***	-0.41***	-0.18***	0.17***	0.19***	0.09***	0.13***
Wikipedia	Outlinks pc	QN	-0.45***	-0.37***	-0.14***	0.19***	0.17***	0.07**	0.10***
OSM	Nodes pc	QN	-0.50***	-0.41***	-0.04*	0.15***	-0.06**	-0.05*	0.01
OSM	Ways pc	QN	-0.51***	-0.41***	-0.07***	0.18***	-0.04	-0.05*	0.03*
OSM	Tags (Nodes/Ways) pc	QN	-0.54***	-0.43***	-0.08***	0.19***	-0.05*	-0.05*	0.01

China

Dataset / Metric			Univariate	Multivariate				
Platform	Attribute	Type	%Urban	%Urban	%Male	%Non-Han	%Age15-64	%College+
Wikipedia	Outlink Entropy	QL	0.28***	0.00	0.04	0.25**	-0.08	0.38***
OSM	Tags per Feature	QL	0.20***	0.10	-0.08	-0.13	0.14	0.04
OSM	% Human Nodes	QL	0.22***	-0.13	-0.10	0.05	0.22**	0.09
OSM	% Human Ways	QL	0.12*	0.10	-0.06	0.13*	0.15	-0.06
Wikipedia	Articles pc	QN	-0.19***	-0.34***	0.16***	0.30***	-0.17*	0.19**
Wikipedia	Length (Bytes) pc	QN	-0.14***	-0.11	0.18***	0.33***	-0.18*	0.36***
Wikipedia	Outlinks pc	QN	-0.13***	-0.29***	0.10**	0.16***	0.04	0.15**
OSM	Nodes pc	QN	0.07	-0.12*	0.17***	0.16***	0.02	0.19***
OSM	Ways pc	QN	0.28***	-0.04	0.14***	0.15***	0.03	0.31***
OSM	Tags (Nodes/Ways) pc	QN	0.17***	-0.09	0.17***	0.15***	0.03	0.26***

Table 4.1 tells a striking high-level story: examining the *% Pop Urban* columns (columns 4 and 5), we see that nearly across the board, *content in peer production repositories about urban areas is significantly different than content about rural areas.* With the exception of the multivariate results for China (a point to which return later), *% Pop Urban* is significant for almost all attributes (quantity and quality) in both repositories and both countries, and its

normalized effect size is often very high. Overall, it appears that peer-produced information about urban and rural areas is of substantially different character.

Below this high-level story there are a number of critical themes that emerge from Table 4.1. The remainder of this section is dedicated to highlighting these themes.

4.5.1. Theme 1: Urban Advantage in Quality

Table 4.1 reveals a strong and pervasive *pro-urban* bias when it comes to our quality metrics (for which the "Type" column = "Quality"). Whether we define quality by manually-assessed Wikipedia quality ratings, content diversity metrics, tag richness, local focus, or human production, *urban peer-produced content appears to be of significantly and substantially higher quality than rural content.* Aside from a few outliers, this result holds for both the univariate and multivariate regressions, across both countries, and across both repositories.

Unpacking the normalized effect sizes into their absolute values, the strength of the urban quality advantage becomes clearer. For instance, for every standard deviation (31.4% absolute) increase in *% Pop Urban*, our univariate regressions indicate that there is a 47.6% relative increase in the percentage of articles in that county that are assessed as C-class or better (41.2% when controlling for sociodemographics through our multivariate regressions).

When considering the percentage of content that comes from editors with a local spatial focus on a given U.S. county (as defined above), we see equally large effects. In purely rural counties (NCHS classification = "6"), *only 4%* of all tokens on Wikipedia pages come from these focused editors, who, as we saw above, contribute nuanced, local information at a much higher rate than "fly-by" editors (and of course bots and batch edits). The equivalent figure for core urban counties (NCHS classification = "1") is over nine times higher at 37.6%. Along the same lines, whereas 4.5% (median) of the tokens in articles about core urban counties

are contributed by bots or batch editors, the equivalent number for entirely rural counties (NCHS code = "6") is 23.4%. In many of these counties, bots and batch editors generated over 60% of their content.

Table 4.1 also reveals a strong urban bias when it comes to content diversity, which is an important metric for both human and machine consumers of peer-produced content. It appears that rural counties have less unique content and more boilerplate information (e.g., census data), limiting the ability of people and machines to determine the unique character of places in these counties.

Interestingly, in many cases the urban bias in quality even persists when controlling for population. For instance, there are far more C-class or better articles *per capita* in urban areas. In other words, there are far more articles that are "useful [at least] to a casual reader" on a per capita basis in urban areas than in rural areas. Our univariate results indicate that with every standard deviation increase in *% Pop Urban* (31.6%), there is a corresponding 24.2% increase in the number of C-class or better articles per capita. We see a similar effect for local tokens: locally-focused editors are contributing fewer tokens on a per-capita basis in rural areas than their urban counterparts, which is likely one cause of the per-capita deficiency of C-class articles. These results point to a systemic underrepresentation of rural editors in Wikipedia, a point to which we return in the discussion section.

It is important to point out the quality deficiencies in rural content are experienced directly by enormous numbers of people (and indirectly experienced through peer-production-based intelligent systems). According to the Wikimedia Foundation's page view statistics (collected via WikiBrain), every month, millions of people visit Wikipedia articles about places in very rural, U.S. counties (and this does not include the people who look at OSM-based maps about these places). Indeed, we aggregated all pages views to all articles about

places in each county over a one-month period, and found that the median, very rural county (NCHS classification = "6") received over 6923 page views (NCHS = "5" counties had a median of 15600). These figures are not a surprise: in the United States, over 46.2 million people live in rural areas [**276**], and many others need information about these areas.

4.5.2. Theme 2: Rural Advantage in Per Capita Quantity

Whereas nearly all of our quality results point to a strong urban advantage, Table 4.1 shows that the opposite is true of the *quantity* of this content. Nearly all of the quantity attributes (Type = "Quantity") in Table 4.1 show a strong and significant negative effect for *% Pop Urban,* indicating a substantial rural advantage in the per capita quantity of peer-produced information.

Examining the *Articles per Capita*, *Nodes per Capita*, and *Ways per Capita* figures, we see that there are indeed many more features of interest in rural areas than in urban areas. For instance, in core urban counties, there is an average of 2,238 potential local editors per article, whereas in purely rural counties this number drops to 467. Given that a miniscule percentage of the populace edits Wikipedia or contributes to OSM [**118**], this places a tremendous burden on local contributors in purely rural counties.

Much of the past work (e.g., [**113, 181**]) that examines content biases implicitly or explicitly assumes that the distribution of content should be roughly equal on a per capita basis, e.g., that a city of, say, 100,000 in sub-Saharan Africa should be described by roughly the same number of articles with roughly the same total length as a city of 100,000 in California. Another important trend in Table 4.1 problematizes this assumption. For instance, consider the United States *Length (Bytes) per Capita* and *Tags (Nodes + Ways) per Capita* rows. Under the "equal per capita" assumption, we would assume that Wikipedia is tremendously

biased towards rural areas, and perhaps that dramatic effort is needed by the community to reduce this bias. However, considering these quantity results in context of the quality results, we see that much of this content is generated by bots, batch editors, and fly-by editors who do not focus in the local area and contribute far less rich local content. Indeed, the end result is that there is actually less content from locally focused editors and the overall proportion of quality articles is less, even on a per capita basis. While raw content is certainly important from some perspectives, this work suggests that it may not tell the whole story.

It is important to note that there is a significant outlier to the general rural advantage in quantity: OSM in China. *In China, there are more ways and tags per capita in urban areas than in rural areas.* A quick examination of the raw data revealed the cause of this outlier: the OSM tools that import massive amounts of spatial data about rural areas in the United States cannot function in China due to government restrictions on the datasets required by these tools. Chinese state law stipulates that geographic data of a certain level of accuracy and scale should be kept secret for national security reasons [**228**]. As such, OSM in China provides a window into urban/rural peer production dynamics when bots and batch editing (e.g., that help to upload government data) do not exist. Without bots and batch editors, not only is the quality of peer-produced content worse in rural areas, but also, in many cases, this content simply does not exist. We return to this issue in the discussion section.

4.5.3. Theme 3: Trends in the Control Variables

In addition to revealing strong, significant, and persistent effects for urban/rural dynamics, our modeling efforts also unexpectedly revealed similar effects for some control variables. Most notably, in the United States, Table 4.1 shows that more Democratic counties and counties with a higher *% White Collar* tend to have higher-quality content. For instance,

unpacking the effect sizes in Table 4.1, we see that for every standard deviation (14.7%) increase in the Obama 2012 vote rate, there is a corresponding 20.6% increase in the percentage of articles that are C-class or better (multivariate model). The control variable results in Table 4.1 provides important motivation for future work that can focus on the corresponding phenomena in more detail.

4.6. Discussion and Implications

The results in this chapter have important implications for peer production communities' efforts to describe rural phenomena, as well as the peer production content generation model more generally.

4.6.1. Systemic Challenges in Describing Rural Areas

Our results point to two distinct challenges for efforts to generate high-quality content about rural phenomena in peer production communities. The first is that rural participation in these communities seems to be lower than urban participation, at least when it comes to contributing content about their home areas. In particular, we observed that there are fewer tokens from locally-focused editors *per capita* in articles about rural areas, with these tokens much more likely to be rich local information than tokens from non-locally focused editors. Reduced participation and its corresponding effects on quality are likely one reason why rural areas have fewer C-class and above articles on a per capita basis than urban areas.

Uneven participation in peer production communities and the corresponding deficiencies in content have been observed in a number of domains, in particular gender. By adapting some of the techniques that have been used to bring women and other underrepresented

groups into peer production editor communities (e.g., [**218**]), this rural participation problem could possibly be partially mitigated.

The second systemic challenge facing peer production communities in describing rural phenomena, however, is much harder to address. Put simply, our results suggest that (1) *it is far more difficult to describe many rural phenomena using peer production than it is urban phenomena* and (2) *the increased difficulty systematically leads to lower quality peer-produced content about rural areas.*

Let us examine this challenge in more detail. Consider, for example, the task of generating a quality article about every incorporated place (e.g., city, town) in the United States. This is a task that Wikipedia has taken on, as the community has determined that all incorporated places are sufficiently notable so as to deserve an article. We know from prior work that a large percentage of contributions to peer production repositories come from locals [**128, 136**], and we also know from prior work [**73, 354**] and the coding study above that local contributions tend to be of higher quality than those of non-locals. As such, for some incorporated places – e.g., New York City – there are literally millions of potential local experts who can help create a high-quality article about the incorporated place. However, for other incorporated places – e.g., Orrtanna, PA (pop. 173) – this number is much smaller. Indeed, we saw above that while there are over 2,200 potential local editors for every Wikipedia article about core urban areas, there are less than 500 in very rural areas.

Given this systemic challenge, it is not a surprise that we found that peer-produced content about rural areas tends to be of much lower quality. Stated more formally, our results suggest the following general property of current models of geographic peer production:

Peer-produced data about rural areas will be of lower quality when the ratio of entities of interest to the size of the local population is much higher in rural areas than in urban areas.

In addition to incorporated places, countless geographic phenomena have more entities-per-capita in rural areas than in cities, and many of these phenomena are being mapped or described in Wikipedia and OSM: e.g., roads, counties, schools, natural phenomena (e.g., creeks, rivers). While not all geographic phenomena display this property – U.S. congressional districts, for instance, are population-controlled – many do. It is reasonable to expect that if rural areas can increase participation rates, they may be able to compete with urban areas on phenomena like congressional districts. However, rural participation rates would have to be orders of magnitude higher than that of urban areas to generate widespread quality content about phenomena for which the number of entities of interest per capita is substantially higher in rural areas (e.g., as with incorporated places).

It is important to note that the basic principle behind the general property of rural peer production stated above is not a new idea. Indeed, in some ways, the general property is an instance of Linus' Law ("Given enough eyeballs, all bugs are shallow" [**255**], which has been explored and characterized in peer production in other contexts (e.g., [**79, 117**]). However, our results highlight the fact that, due to the inherent, low-population-density nature of rural areas, peer production communities will in general find themselves on the wrong end of Linus' Law when trying to describe or map rural phenomena.

4.6.2. Different Model of Peer Production for Rural Areas

Automated and partially-automated software agents that generate content can be quite controversial in peer production communities [**194, 355**]. However, our China OpenStreetMap

results show what happens to rural peer-produced content without these agents: not only is content of lower quality, but also, in many cases, it simply does not exist. While relying solely on high-quality manual edits may be possible for content about urban areas, our research demonstrates that this is not true for rural areas.

More generally, our results point to descriptions of rural areas benefiting from a different model of peer production than exists in cities. In this model, bots and batch editing play a more central role to partially account for the reduced amount of local expertise per entity of interest. By embracing this peer production model and developing new technologies to support it, it may be possible to increase the quality of content about rural areas, especially as automated content generation technologies become more effective.

For example, new tools (e.g., Reasonator [**204**]) are becoming available that attempt to turn information from Wikidata, Wikipedia's structured data sister project, into natural language Wikipedia articles. Wikidata supports information well beyond that available in government sources like a census, allowing Reasonator tools to generate text that more resembles that contributed by editors with a local focus (e.g., information about a town's mayor or its prominent citizens). The Wikipedia community has been heavily critical of incorporating content from Reasonator and related technologies, but content about rural areas may benefit significantly from this content in the future. Additionally, because quality in rural areas is already low, rural articles provide a useful do-no-harm (or do-little-harm) testing ground for these technologies.

Similar approaches in OSM are also possible. For example, automated approaches could be developed to extract semantic information (e.g., opening hours) from business' websites and assign this information as tags to the corresponding OSM entity. Computer vision operating on satellite imagery may also help increase rural data quality.

4.7. Limitations and Future Work

The research presented above presents a number of opportunities for follow-on work:

(1) While this research examined an Eastern country and a Western country, our results should be confirmed in a variety of other cultural contexts as well. Moreover, a more qualitative investigation of the complex rural/urban editing choices being made in each online community and in each cultural context would be quite informative.

(2) It may be useful to consider our findings in the context of long-standing discussions about whether a particular class of entity is sufficiently notable for Wikipedia articles. When the entity class under consideration is geographic in nature, tools based on our work can inform this debate by predicting expected quality in urban and rural areas.

(3) An exciting area of future work arises out of the possibility of encouraging urban contributors to "adopt" a rural region, learn about that region, and become local specialist editors in that region, even if they do not live there (although the effectiveness of this approach would have to be measured carefully).

4.8. Conclusion

Examining both Wikipedia and OpenStreetMap in both China and the United States, this research showed that peer production faces systemic challenges in describing rural phenomena, challenges that will persist even if the observed participation issues are addressed. More generally, this work adds to a growing body of literature that suggests that urban/rural dynamics play a key role in geographically-referenced content that is produced in online social systems. We hope our results encourage further investigation of these dynamics, as well as the development of tools and strategies to help mitigate the identified problems.

CHAPTER 5

Geolocation and Algorithmic Bias

In this chapter, we explore how population bias in Twitter affects geolocation inference algorithms trained from this data. Specifically, we focus on the urban-rural divide and consider whether rebalancing of training data is sufficient to build "fairer" algorithms.[1]

5.1. Introduction

As social media adoption has increased dramatically, research that takes advantage of this publicly-available, real-time stream of information has also become quite prevalent. In addition to facilitating new discoveries about human behavior in the social sciences and related fields (e.g., [**4, 55, 56**]), social media has also been a major catalyst in the development of new intelligent algorithms. For instance, researchers have used social media to develop new recommender systems (e.g., [**185, 285**]), summarize the character of cities (e.g., [**45, 346**]), and infer the location of Internet users (e.g., [**161, 191, 211**]). The ability to leverage this content as training data for algorithms stands apart from its quality for users or appropriateness for research (as discussed in Chapters 3 and 4).

Recently, concerns about population bias in social media have been the subject of much discussion (see §2.3.1). Social media population bias, or the notion that different demographic groups may participate in social media platforms at different rates, has been found

[1]The work presented in this chapter was originally published in: **Johnson, I.**, McMahon, C., Schning, J., and Hecht, B. The Effect of Population and "Structural" Biases on Social Media-based Algorithms – A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. *ACM Conference on Human Factors in Computing Systems 2017*.

to be endemic to most social media datasets. Much work has gone into quantifying and understanding these biases, with significant biases being found along gender (e.g., [**239**]), age (e.g., [**187**]), race (e.g., [**222**]), income (e.g., [**201**]), education (e.g., [**187**]), and urban/rural lines (e.g., [**140**]).

Many researchers who use social media to understand human behavior have recognized the importance of correcting for social media population biases in their studies (e.g., [**57, 265**]). They have also begun the critical work of developing best practices for doing so (e.g., [**265**]). However, the same is not true in literature on social media-based algorithms: even though it has been hypothesized for several years that population bias would affect social media-based algorithms (e.g., [**140, 222**]), little research has been done to investigate this hypothesis (let alone identify remediating measures).

The goal of this chapter is to help address this gap in the literature. Since the space of social media-based algorithms and social media population biases is extensive, we initiated our exploration with a focused but important case study: examining the performance of Twitter geolocation inference algorithms across the U.S. urban-rural spectrum. Due to the rising import of geographic information, geolocation inference algorithms – usually focusing on Twitter data – have attracted widespread interest across computer science and related fields (e.g., [**68, 125, 160, 211, 247, 50**]). The aim of these algorithms is to predict the location of a Twitter user or her tweets using implicit signals. This is typically done by examining the content of the user's tweets and/or the geographic configuration of her explicitly encoded social ties.

Similarly, U.S. rural-urban biases in social media have also attracted a growing amount of attention in the literature (e.g., [**105, 140, 157, 351**]). This is likely in part due to the relatively large effect sizes involved: people who live in rural areas often participate in

social media at a fraction of the rate of their urban counterparts and contribute orders of magnitude less content. Twitter is no exception to this trend (e.g., [**157, 293, 351**]). The urban-rural population-bias effect sizes are of particular interest given the significant size of the rural population: over 46 million people live in rural counties in the United States (close to double the population of 18-24 year-olds) [**179**].

Rather than focusing on a single geolocation algorithm to assess the effect of urban/rural population bias on algorithm effectiveness, we study two well-known geolocation inference algorithms: that of Priedhorsky et al. [**247**] and that of Jurgens [**160**]. We chose these two algorithms because they span the range geolocation inference algorithm design, allowing us to more robustly evaluate the effect of population bias. Priedhorsky et al.'s and Jurgens' algorithms fundamentally differ in two key ways: methodological paradigm and problem definition. With respect to paradigm, Priedhorsky et al.'s algorithm is representative of text-based approaches to geolocation inference (e.g., [**40, 138, 183, 200, 258**]) while Jurgens' is representative of network-based approaches (e.g., [**13, 50, 211**]). Similarly, Priedhorsky et al.'s algorithm seeks to solve the "geolocate a tweet" version of the problem, while Jurgens' algorithm addresses the "geolocate a user" version of the problem.

As we will describe below, we find that regardless of methodological paradigm or problem definition, geolocation algorithms underperform for rural users. In some cases, the effect sizes are dramatic. For instance, the text-based algorithm is able to geolocate urban tweets within 100km of their true location at a rate 2.3x greater than is the case for rural tweets.

However, a major result in this chapter is that population bias is not the only driver of population-variable accuracy in social media-based algorithms. Rather, we find evidence that design choices in social media-based algorithms can also have a powerful biasing effect. That is, our results suggest that algorithmic bias is a function to a large degree of both

population bias in the underlying social media dataset and *structural bias* that arises from algorithmic design choices. We also observe early evidence that text-based algorithms may be more liable to structural bias than network-based algorithms. In particular, we found that when population bias was removed through balanced resampling and oversampling, our network-based algorithm showed much improved performance on rural users, but the same methods were less effective in our text-based algorithm.

This chapter has several implications for those who develop and study social media-based algorithms. For instance, our research provides additional weight to calls (e.g., [**267**]) to consider the design of algorithms rather than just the output of these algorithms when evaluating intelligent technologies for bias. In other words, our results directly motivate further research on algorithmic design decisions that avoid structural bias. Moreover, as we will discuss, our results point to specific solutions to structural bias in the geographic algorithm domain. Similarly, our results also highlight the dangers of global evaluation metrics for social media-based algorithms, providing a data point that shows that these metrics can hide poor performance for certain populations and establish perverse incentives to reduce performance for these populations even further when they are minorities.

In addition to showing that both population and structural bias can result in uneven performance by social media algorithms, this research also makes more specific contributions to the large body of research on geolocation inference algorithms. Namely, by demonstrating that these algorithms' performance is geographically variable in a systemic fashion, we show that it is likely that these algorithms have introduced additional bias into the studies and systems that have used their output. This raises the stakes for quickly identifying solutions and establishing best practices. It also suggests that researchers who use these algorithms

should be careful to audit their output as we have done here prior to incorporating them into their larger systems.

Finally, it is important to note before continuing that while we find important biases that are robust across two separate algorithms, this is a case study on Twitter geolocation inference rather than an exhaustive survey of population bias's effect on all social media-based algorithms. Given their import, identifying potential biases in social media-based algorithms is a serious matter, as is developing means of addressing them. However, because our findings our limited to a single algorithm family (albeit an important one), our findings should be interpreted as an important preliminary step in these directions rather than the definitive answer to the associated questions.

5.2. Related Work

This work builds on research in three main areas: 1) characterization of population bias in social media, 2) social media-based geolocation algorithms, and 3) the literature on algorithmic accountability. We discuss these areas and their relationship to our work below.

5.2.1. Population Bias in Social Media

Population bias is a well-studied issue in social media. Since 2005, the Pew Research Center has conducted annual surveys of social media usage in the United States. These surveys [**239**] show that social media participation rates vary extensively across gender, race, education, socioeconomic status, and urban-rural lines.

Augmenting the Pew findings, many researchers have investigated this problem by analyzing the geographic distribution of posts across social media sites and making demographic

inferences from census data or from users' self-reported information. The demographic dimension of analysis we choose for our study – the urban-rural spectrum – has been the focus of some of this work. For instance, Hecht and Stephens [**140**] studied Foursquare, Flickr, and Twitter and found a consistent pro-urban bias (e.g., there are 3.5 times more Twitter users per capita in urban areas than rural areas). Similarly, Malik et al. [**201**] demonstrated that there are higher densities of tweets in urban, young, and rich areas. Gilbert, Karahalios, and Sandvig [**105**] found substantial differences in behavior between urban and rural users in Myspace. Many other papers have studied population bias issues across other geographies, platforms, and demographics, and, to our best knowledge, some form of population bias has consistently appeared in all of this work.

A few researchers have attempted to explicitly account for population bias in their own studies. Culotta [**57**] demonstrated that tweet-based models of public health indicators saw improved performance when their Twitter dataset was balanced for race and gender by county. Landeiro and Culotta [**183**] examined how to control for shifts in the magnitude of population bias within input data to classification algorithms. Finally, Pavalanathan and Eisenstein [**237**] have explored how people of different ages and genders tweet differently and how this relates to performance in geolinguistic algorithms. They balanced their tweet samples based on the total population by county but did not find that this significantly impacted their algorithm.

We are motivated by the above work and the questions that it raises. While Culotta saw improved precision when controlling for specific population biases, Pavalanathan and Eisenstein did not when controlling for more general population density, but still found vast disparities in performance of the algorithm across age and gender. These divergent

results raise the question of how algorithmic bias arises and is manifest in social media-based algorithms as well as the relationship between algorithmic bias and population bias in social media. Our results explain the conflicting findings in this motivating work by delineating the concept of structural bias and showing how it can counteract or exacerbate population bias. Moreover, by also examining network-based algorithms in our research, we are able to speak more broadly to social media-based algorithms in general and not just a single class of algorithms.

5.2.2. Geolocation Inference Algorithms

A large portion of the research and applications associated with Twitter data has a geographic component. However, this research is limited by the fact that only 1-2% of tweets are geolocated [**50**]. As a result, geolocation inference algorithms for Twitter, which attempt to uncover tweet and user locations that have not been explicitly disclosed, have become a very common direction of study.

There are two main classes of Twitter geolocation inference algorithm (see [**161**] for an overview): text-based, which predict the location of a tweet based on its content, and network-based, which predict a home location for a user based on their connections to other users.

Text-based geolocation algorithms rely on the tendency for language usage to vary as a function of geography (e.g., [**74**]). By extracting lexical features and concepts local to an area from the text during a training phase (e.g., sports teams, regional vernacular, a town name), these algorithms can build models that predict the geographic location of a new tweet based on its content. These algorithms generally either attempt to model text features as

an explicitly spatial process (e.g., [**247, 258, 323**]) or treat the geolocation problem as a classification problem among administrative units (e.g., cities) (e.g., [**44, 125, 200**]).

Network-based geolocation algorithms rely on the social network in which social media posts are typically embedded. In these algorithms, explicitly encoded network ties or user interactions are used to build an egocentric social network for the user whose location is desired. Any known locations of the user's neighbors are combined to predict the location of the user [**161**]. This approach leverages a fundamental principle in human geography that, in general, interaction decreases with distance (e.g., [**116**]), meaning that connected users are likely close geographically [**13**].

Though our goal is to probe algorithmic bias within social media-based algorithms in general, selecting geolocation algorithms as our case study has two additional benefits: both text-based (e.g., [**40, 138, 183, 200, 247, 258**]) and network-based (e.g., [**13, 50, 160, 211**]) geolocation inference have been a major area of interest in the past few years, and our work can help lead to more equal and effective approaches in the future. Second, because of this robust literature on both text- and network-based approaches, we are able to explore bias in social-media based algorithms that draw on two of the most prominent methodological paradigms, improving generalizability and affording cross-paradigmatic comparisons.

5.2.3. Algorithmic Accountability

Our research builds on the growing literature on algorithmic accountability (e.g., [**9, 268, 267**]), in which algorithms are probed for discrimination or other societally undesirable outcomes. Our work extends the accountability literature to include well-known geolocation inference algorithms (and the rural-urban divide). Additionally, much of the algorithmic

accountability literature has focused on detecting algorithmic bias when faced with a black-box system (e.g. [**41, 301, 166, 291**]). Our research focuses on algorithms with a published description, open-source code, and accessible data, enabling us to investigate biases at a more detailed level, determine the potential causes of these biases with more certainty, and begin to learn how to mitigate these biases. In the discussion section, we highlight how our findings related to structural bias emphasize the importance of lower-level analyses (as per Sandvig et al. [**267**]) and open-source implementations in understanding and reducing algorithmic bias.

5.3. Methods and Data

In this section, we describe our two focus geolocation inference algorithms and the datasets they use in more detail. In general, our approach to working with these algorithms was to replicate the choices and approach taken in the corresponding papers. In the few cases when this was not possible, we deferred to other best practices in the geolocation literature as is explained below.

5.3.1. Text-based Algorithm

We selected the text-based algorithm from Priedhorsky et al. [**247**] for our analysis. We chose this algorithm because it is representative of many text-based algorithms in its general approach (e.g., it calculates a geographic layer for each term and utilizes a standard set of Twitter text and metadata fields), has made an impact since it was published (e.g., [**125**]), and its source code has been made available by its authors[2], which minimizes the risk of implementation error and provides us with full control over the algorithm and its inputs.

[2]https://github.com/reidpr/QUAC

The algorithm is trained on a set number of tweets with known locations and builds Gaussian mixture models (GMMs) for tokens in the text of the tweet, its user's time-zone, self-reported location field, and specified language. The GMMs capture the probability that a given token originated from an area based on the training data. A prediction for a given tweet is made by tokenizing it, weighting and combining the individual GMMs for each token in the tweet, and making a prediction based on the highest probability area in the resulting GMM.

5.3.2. Network-based Algorithm

We selected the algorithm by Jurgens [**160**] for our network-based algorithm. We chose this algorithm because its performance is in line with other state-of-the-art approaches [**161**]. Additionally, like Priedhorsky et al.'s algorithm, code for Jurgens' algorithm has been provided by the author[3], minimizing implementation risk and allowing for direct manipulation.

Jurgens' algorithm builds a bi-directional mention network by generating an edge between two users only when both users have mentioned each other in a tweet. A mention occurs when a user includes another user's username in a tweet using the "@" notation (e.g., "President @barackobama will be speaking tonight"). Starting with a training set of users with known locations, Jurgens' algorithm iteratively propagates the location of the known users to any of their neighbors who have not been successfully located, inferring the location of a newly located user in each iteration as the median of their previously located neighbors. Jurgens repeats this process for five iterations and we do the same in our analyses.

[3]https://github.com/networkdynamics/geoinference

5.3.3. Social Media Datasets

We built our tweet dataset for our text-based algorithm following standard practices in the text-based geolocation inference community (e.g. [**40, 74, 183**]). More specifically, we utilized the Twitter Streaming API with a bounding box configured to the contiguous United States, which, like many geographic studies of geotagged content in the U.S. (e.g. [**187, 223**])[4], was the geographic extent of our analyses. In line with prior work (e.g. [**40, 74, 183**]), we left open our tweet collector for one month. Only tweets with coordinates in the contiguous United States were retained, resulting in a dataset of 51.2 million tweets from 1.6 million unique users from October 2014. All of these tweets are explicitly geotagged with the latitude-longitude coordinates from where the tweet originated, which is necessary to provide ground truth of the algorithm. One important exception to our approach relative to that of Priedhorksy et al. was that we used geographically bounded version of the Streaming API rather than the random Streaming API. We made this exception for a simple reason: we required a much larger dataset in our study region in order to have sufficient data in rural areas for our experiments.

Network-based approaches utilize different techniques to collect data than text models and, as such, it was important to collect a separate dataset to be in accordance with standard practices in the network-based domain. To gather data for our network model, we adopted the methodology used by Jurgens et al. [**161**], which involved building a mention network from a dataset of randomly collected tweets (our dataset started with 99 million tweets from 26 million users from August and September 2015). We further restricted this dataset to only consider tweets from users we could geolocate to the United States, which narrowed our

[4]Geographic methods often assume a contiguous region, and this is the case for the methods we employ here. Moreover, given their populations, it is highly unlikely that the inclusion of Alaska and Hawaii would have significant altered our results.

dataset down to 3.2 million tweets from 1.2 million users as described below, of which 113K comprised the ground truth of our final mention network (see below for more information about ground truth development). Though this is smaller than that used by Jurgens et al., it is in line with other network-based algorithm studies [**161**].

5.3.4. Ground Truth Data

As noted above, we selected the urban-rural divide as our focus demographic dimension because it corresponds to a well-studied population bias that exists in most forms of social media [**140**]. To categorize locations along the urban/rural spectrum for our second demographic variable, we follow standard practice in the literature that looks at urban/rural issues in online communities [**140**]. Specifically, we utilize the U.S. National Center for Health Statistics' Urban-Rural Classification Scheme for Counties [**151**], which assigns each county in the United States an ordinal code from 1 (most urban) to 6 (most rural).

Using these codes, it is straightforward to obtain urban/rural data for our text-based algorithm's ground-truth dataset. Namely, since this dataset consists entirely of geotagged tweets in the contiguous United States, we simply perform a reverse geocoding operation that labels each tweet with the county in which it is located. We then assign each county's urban/rural code to the tweet.

While text-based algorithms predict the location of each tweet individually, as noted above, network-based approaches typically seek to locate a given user (and assign all of that user's tweets to that "home" location). There are two methods used in the literature for determining a user's home location for use as ground truth in these algorithms (see Johnson et al. [**157**] for a complete overview of ground truth identification techniques). Jurgens [**160**] and McGee, Caverlee, and Cheng [**211**] respectively define a user's home location as the

geometric median of their tweets given a minimum of five (three) geotagged tweets within 30 kilometers (50 miles) of each other. The other method takes advantage of the user's self-reported location field. This method has the drawback of being quite noisy [**138**] and though Jurgens et al. [**161**] found the location field to lead to lower overall precision, it is a method that has been used routinely in the literature [**188**].

We first evaluated the geometric median method for our dataset but found the resultant dataset to be prohibitively small, with only 20,000 users and a miniscule number of rural users. As such, we turned instead to geocoding the location field through an approach that has been shown to significant reduce the noise in this data source (albeit with reduced recall) [**137**]: using a Wikipedia-based geocoder. We leveraged the geocoder in WikiBrain [**273**], which resulted in a much larger dataset of 1.2 million users from which we built our mention network.

In line with previous results on population bias, we found that, relative to our census data, the most urban users (NCHS class 1) were highly overrepresented in our datasets (130% and 210% relative to their proportion in the overall population for the text-based and network-based datasets respectively). The most rural users (NCHS class 6) were accordingly underrepresented (45% and 24% relative proportion for the text-based and network-based datasets respectively). Lastly, it is important to note that rural/urban labels were used for evaluation purposes only and are not provided to the algorithm with training or testing data.

5.3.5. Evaluation Framework

When evaluating the two algorithms, we followed the procedure outlined by the two corresponding papers as closely as possible to measure the bias present in typical performance. We use five-fold validation for each network-based model. For the text model, we build and

test five models for each condition. We constrain the test tweets in each text-based model to those from the days following the dates of the tweets that comprise the training data, ensure no overlap in users between training and testing phases, and use training set sizes (30,000) equal to those used by Priedhorsky et al. However, because we were limited to the Streaming API rather than the higher-volume "gardenhose" API for collecting our Twitter data, we limited the size of our network model training datasets to 24,000 users (selected anew for each fold out of the 113,000 who comprised our ground truth data), which is smaller than that used by Jurgens. We examined training datasets up to 40,000 for our baseline model though and found similar trends.

In all cases, we define algorithm precision as the percentage of predictions within 100km of the actual location. We tested different values for the distance (20km, 50km, 200km, 500km) as well as defining a true positive as a prediction lying within the same county as the ground truth (much research aggregates tweets to counties in order to leverage census data). All of these definitions led to similar patterns of results. We also tested the text-based algorithm with a similar dataset of tweets from June 2015 and saw similar trends. We only report recall (the percentage of users or tweets for which an algorithm could provide a location) for the network-based models because the recall was consistently around 100% for the text-based models.

In line with the NCHS classifications [**151**], we combine data from counties with NCHS codes of 1 and 2 ("large metropolitan counties") into one "urban" class and counties from NCHS codes of 5 and 6 ("nonmetropolitan counties") into one "rural" class. We use these definitions of urban and rural for the rest of the chapter. We restrict our reporting and discussion to just these urban and rural classes, though the results for NCHS categories 3 and 4 generally fell between those for the urban and rural classes.

5.4. Results

We structured our exploration of the effect of population bias on social media-based algorithms by asking three cascading research questions. Our first question was whether the two geolocation algorithms exhibit biased performance in favor of urban areas (**RQ1: Is there an algorithmic bias in the direction of the population bias?).** Our next step was to inquire whether any identified bias was due to population bias (**RQ2: Is any bias due to population bias?**). We did this by examining the change in algorithmic bias when we adjusted the training dataset so that it contained a representative urban/rural sample of the general population.

Finally, given a positive answer to RQ2, we planned a third research question whose objective was to investigate whether any remaining bias could be eliminated by training solely on the algorithmically disadvantaged population (**RQ3: Can any remaining underperformance for a specific population be fixed by training solely on data from that population?**). This is equivalent to building a separate algorithm customized (trained) specifically for rural users and tweets and another for urban users and tweets. The goal of RQ3 was to identify whether any bias that remained after adjusting for any population imbalances was inherent to the algorithm.

Below, we use our three research questions to structure our discussion of our results. To determine the significance of changes in precision of the algorithm, we adopt the methods used by Compton et al. [**50**], which set confidence intervals as the average precision ± 1.5 times the interquartile range for the folds. Comparisons are only made when confidence intervals do not overlap unless otherwise noted.

Table 5.1. Geolocation precision by urban / rural.

Results for typical training data as well as various population bias manipulations.
Precision: confidence intervals for percentage of predictions within 100 km of true location.
Recall: values in parentheses; the text-based models had recall always around 100%.
*The network-based rural-boosted model was trained on 6000 tweets due to lack of data.

	% of Training Data		Precision (Recall)		
Text-Based Models	Urban	Rural	Urban	Rural	Overall
Typical (Baseline)	63.80%	9.00%	23.0 ± 3.0%	9.9 ± 0.4%	20.6 ± 1.9%
Population Bias Balanced	55.30%	15.00%	22.1 ± 1.5%	10.5 ± 0.6%	20.4 ± 1.1%
Urban Boosted	100%	0%	27.6 ± 3.0%	4.9 ± 0.3%	19.5 ± 2.0%
Rural Boosted	0%	100%	5.0 ± 0.5%	17.9 ± 1.5%	6.8 ± 0.6%
Network-Based Models	Urban	Rural	Urban	Rural	Overall
Typical (Baseline)	75.00%	5.20%	25.7 ± 0.4% (13.1%)	20.6 ± 1.7% (8.1%)	25.0 ± 0.4% (12.3%)
Population Bias Balanced	55.30%	15.00%	20.6 ± 0.8% (12.4%)	39.2 ± 5.2% (9.5%)	22.2 ± 0.8% (11.9%)
Urban Boosted	100%	0%	27.0 ± 3.9% (13.3%)	5.0 ± 1.9% (9.0%)	22.3 ± 3.2% (11.6%)
Rural Boosted*	0%	100%	1.0 ± 0.3% (4.6%)	59.2 ± 5.3% (8.4%)	3.8 ± 0.4% (4.5%)

5.4.1. RQ1: Is there algorithmic bias?

Examining the precision of each algorithm in urban and rural contexts (Table 5.1, rows labeled "Typical (Baseline)"), a clear pattern of bias emerges. Both algorithms perform worse for rural users than for urban users, with the magnitude of the bias being greatest in the text-based algorithm: this algorithm is able to accurately locate urban users within 100km at a rate approximately 2.3 times greater than that for rural users. The equivalent precision number for the network model is 1.3x, and recall is 1.6x better for urban users than rural users in the network model as well.

Figure 5.1 shows the precision of the text-based algorithm by county in the contiguous United States. It demonstrates the depth of this bias – almost all high precision clusters center in urban areas around cities.

In addition to motivating further inquiry as to whether this algorithmic bias arises from population bias in the underlying dataset or other factors (i.e. our RQ2 and RQ3), this result has important implications in and of itself. Namely, geolocation inference algorithms have served as inputs to systems and studies and our results establish for the first time that

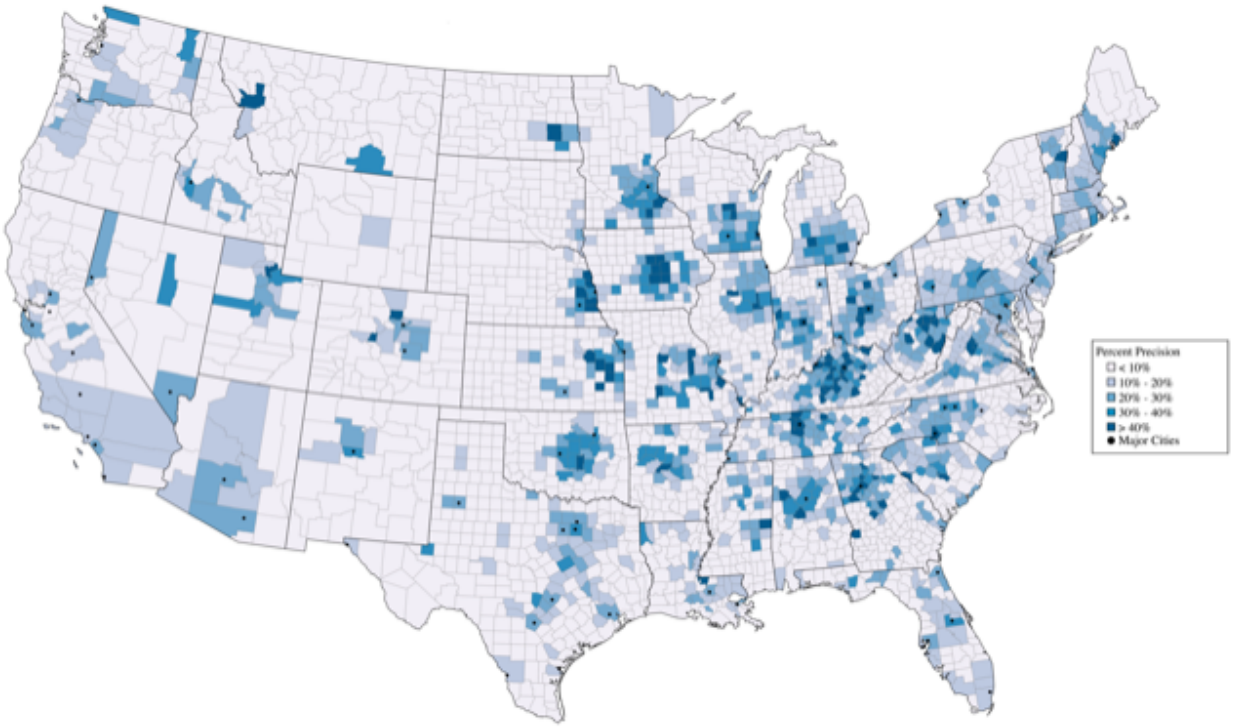


Figure 5.1. Map of text-based geolocation precision.
Text-based geolocation baseline precision by county (percentage of tweets originating from each county correctly geolocated to within 100km of each tweets true location).

the use of these well-known geolocation inference algorithms from the literature will inject further population bias into any geographically referenced dataset of Twitter users. For instance, in the case of the text-based algorithm, 2.3x more urban users than rural users will be "put on the map" correctly. We return to this point in the discussion section.

5.4.2. RQ2: Is the observed bias due to population bias?

The cells in the "Typical (Baseline)" rows and "% of Training Data" column in Table 5.1 indicate that, as noted above and in line with a number of previous studies on Twitter, our unadjusted training data has significant underlying population biases. For instance, we can see that 63.8% of our text-based urban/rural ground-truth dataset can be classified as urban

according to our definition, but only 9.0% of our dataset can be classified as rural. The actual census proportions would be 55.3% and 15.0% respectively, indicating a strong urban bias in our dataset.

Our goal with this research question was to determine whether correcting for this population bias in the training datasets is the primary cause of the biases we observed in our RQ1 analyses. As pointed out by Pavalanathan and Eisenstein [**237**] and Burger et al. [**36**], more data about a group in prediction tasks generally improves accuracy of the algorithm. Therefore, we expect that by adjusting for population bias, we can induce greater and perhaps equal performance for underrepresented populations (i.e. rural populations). Though rural areas still would have a much lower number of tweets even within a population-balanced dataset, maintaining a consistent number of tweets per person in an area as training data should capture an equal percentage of an area's location-indicative words (to use Han, Cook, and Baldwin's [**125**] vocabulary).

To answer our second research question, we performed a simple modification of the datasets on which we trained each algorithm: instead of resampling the datasets at random for each fold of training the algorithm as we did in RQ1, we resampled them such that they were representative of the demographics of the general population with respect to the rural/urban divide. In other words, we generated training datasets without population bias. We did not adjust how we sampled data for testing each fold (i.e. it remained random).

The results of the evaluation of each algorithm under this population-balanced condition can be found in Table 5.1 in the rows labeled "Population Bias Balanced". These rows tell a relatively straightforward story: while using a representative sample of the general population significantly changed the results for the network-based model with respect to

the urban/rural divide, balancing the population had only a negligible effect on the text-based model. In other words, the text-based model remained significantly urban-biased, even when using a training set that removes population bias in the underlying Twitter dataset. On the other hand, the network-based model was significantly less biased when trained on a population-balanced dataset.

Like was the case for RQ1, this result both motivates the investigation of our subsequent research question (RQ3) and has implications on its own. Whereas there is evidence that addressing the effects of demographic bias in social media-based social science research can be done by simply resampling the underlying dataset [**57**], our findings suggest the same cannot be said for social media-based algorithms, at least in the case of our text-based model. This is another subject to which we attend below.

5.4.3. RQ3: Can we fix biases through oversampling?

Even when the representation of rural Twitter users has been boosted to match that of the general population, rural users still make up a much smaller relative proportion of the training set and therefore have a much smaller corresponding absolute number of training samples. In this experiment, we sought to see if this imbalance in absolute number of tweets can explain the algorithmic bias we observed above.

We did this by training and testing on each demographic group separately (i.e. separate models for rural and for urban). If separate models perform equally well (e.g., rural precision is as high as urban precision), then we would know that the algorithmic bias in our algorithm arises solely from population characteristics of the input dataset. If bias still exists even in separate models, then we would know that there are structural biases within the algorithms

that prevent equal performance for these demographics, no matter their representation in the training dataset.

The results of this experiment can be seen in the rows labeled "Urban Boosted" and "Rural Boosted" in Table 5.1. Of particular note is the performance of the text-based algorithm. Even when training and testing solely on tweets from rural users, the text-based model cannot geolocate these tweets as well as it can for tweets from urban users when using a simple random sample or population-adjusted training set (let alone using an only-urban model). While we do see a large and significant improvement in the rural case, performance still falls short of that of urban tweets for all of our models that contained any urban training data at all. We also note that we tried boosting the absolute number of training samples by up to two times the number used in Priedhorsky et al., and the results were consistent: the rural-only model still had a precision less than the random and population-adjusted urban precision. In other words, no matter the training data, there appears to be something in the design of the text-based algorithm that prevents it from performing well for rural users relative to urban users.

The story for the network-based model is different, with it appearing to also suffer from structural biases, but not to the same degree as the text-based model. With respect to precision, the model trained only on rural users actually outperforms the model trained only on urban users. However, the recall remains lower than any of the urban models.

These results have important implications for the design and application of social media-based algorithms. We begin our Discussion section below by highlighting these implications.

5.5. Discussion

5.5.1. Algorithmic Bias = Population Bias + Structural Bias + ϵ

The results to our second and third research questions suggest a nuanced understanding of the mechanisms behind algorithmic bias. Namely, while we saw that some algorithmic bias could be explained by population bias in the underlying training sets (e.g., the gap in performance between urban and rural users in the network-based algorithm), not all of it could. In fact, in the case of the text-based algorithm, even dramatically overcorrecting for population bias by training solely on rural users did not make the algorithm perform as well as it typically does for urban users.

These findings support the notion that algorithmic bias must be understood as a function of both population bias and structural bias inherent to the algorithm's design (as well as other factors that have yet to be discovered). In other words, there is something about the nature of some algorithms that inherently biases them towards lower performance for certain populations.

5.5.2. A Closer Look at Structural Bias

Examining the structure of the two algorithm families under consideration, a number of hypotheses emerge for their differing amounts of structural bias along urban/rural lines. Network-based algorithms build an egocentric network, so a prediction for a given user is affected directly by her/his social network neighbors and potentially indirectly by other nearby users (e.g. neighbors of neighbors) through multiple stages of inference. This means that addition or subtraction of users in one part of the network largely does not impact a given user elsewhere in the network even though it can boost precision and recall amongst

their more immediate social network neighbors. Indeed, within our mention network, we find very high homophily: 90% of edges between users with known ground-truth locations are of the same type (i.e. for a user known to be in an rural class 6 county, there is a 90% chance that any of their social network neighbors whose location is known a priori are also in rural class 6 counties) Boosting data for rural users therefore greatly increases the likelihood that a rural user in the testing dataset will have neighbors that are located without affecting most urban users because they would not be closely linked through the network to the rural users.

Text-based algorithms, on the other hand, see much greater dependencies between users. Toponyms and language that have broad usage across the country will be skewed towards being located in urban areas with their higher density of users. Furthermore, when examining the average number of words per tweet that can be identified as geographic Wikipedia concepts (i.e., words tied directly to place) through wikification algorithms implemented by Sen et al. [**273**], we see that the most urban tweets (NCHS code = "1") have 25% more "wikifiable" words per tweet than the most rural tweets (NCHS code = "6"). Since location-specific words such as these are key to how the text-based algorithm operates, some of the urban advantage may come from differing language patterns and topics of conversation in tweets across urban-rural lines [**74**].

Another common design decision in text-based algorithms that could be a cause of structural urban/rural bias relates to the fixed distance parameters central to many of these algorithms' low-level functionality. These distance parameters, which manifest as grid-cell sizes or geographic probability distribution ranges, fail to take into account the varying distances at which the use of a term is predictive in urban versus rural areas. For instance, when someone tweets a name of a high school mascot in an urban area, that name is predictive

for a smaller area than the same situation in a rural area (i.e. rural areas have significantly larger school districts by area). This problem can be understood in geostatistics terms: the use of a fixed-distance parameter assumes a fixed range of spatial autocorrelation for tweet usage, which likely is not true across urban/rural lines. In more general terms, this means that employing fixed distance parameters will fail to capture the full predictive power of each rural tweet in text-based algorithms, and rural areas already have fewer tweets per capita to begin with.

5.5.3. Achieving Parity

Though lower overall recall and precision is a barrier to implementation of network-based algorithms, our work indicates that it may be easier to achieve parity within network-based algorithms by boosting data collection efforts around underrepresented populations. Although more work should be done to confirm this effect in other types of social media algorithms, researchers and designers for whom equity is a top priority may want to consider utilizing network-based methods when doing Twitter-based geolocation.

It is also important to note that the contribution of structural bias to algorithmic bias we have identified here adds weight to the argument of Sandvig et al. [**267**] and others that algorithmic accountability work needs to consider algorithms at levels deeper than simply inputs and outputs and that algorithmic accountability research teams have people with "algorithmic skills that allow a facility with the relevant [algorithmic] ideas" [**267**]. These skills will be necessary to identify and address bias at the structural level.

Similarly, the notion of structural bias highlights the importance of open-source implementations of algorithms. Open-source affords the ability to understand and design fixes

for these algorithmic biases because we are able to examine and manipulate how these algorithms function. Our results show that we cannot rely solely on adjusting the data going into the algorithm to achieve parity.

5.5.4. A Tradeoff Between Equity and Effectiveness

There is one column in Table 5.1 that we have yet to discuss in detail: the "Overall" column. This column indicates the performance on the entire randomized population (i.e. data in an unmodified proportion of users and tweets). In other words, the "Overall" column reports the precision (and recall) a researcher or developer would expect to achieve if she were to apply the model listed in each row to all of Twitter.

There is a clear trend in the "Overall" column: for models in which rural performance is better, the performance on overall population gets worse. From the perspective of a researcher or developer, this means that in order to improve rural accuracy, one has to reduce overall accuracy. Furthermore, for the text-based algorithm, we also tested a wider variety of urban-rural training data proportions to understand the responsiveness of the algorithm to smaller shifts in data. In doing so, we found that peak performance (21.0% overall precision at 100 km) came in a model that removed a quarter of the rural training data and boosted urban accordingly. Rural users performed very poorly in this model (6.8% precision), but the slight increase in precision for urban users (23.5% precision) along with their inherently higher proportions in the testing data was sufficient to boost the overall precision.

Our results demonstrate that, at least in the case of Twitter geolocation, that there is a clear trade-off between equity and effectiveness, a result for which there is evidence in other algorithmic accountability contexts as well (e.g., [**166, 234**]). An important corollary to this

tradeoff is that using single measures of precision and recall for an algorithm can gloss over very real, non-random variation in algorithmic performance for different groups of people. Until better algorithms can be developed that do not compromise equity (or equality) for effectiveness (e.g., by addressing structural bias), algorithm designers should conduct and report a more thorough examination of performance across different populations as has been advocated by many in the algorithmic accountability community, especially when past work has suggested that there may be population bias or structural bias.

5.5.5. Privacy

We have conducted this research with the assumption that higher precision and recall is a desirable outcome for a given population. While this is a valid assumption in an application of geolocation like public health tracking, this is not always the case for social media-based algorithms and for geolocation specifically. It is important to note that in this light, there may be benefits to being "disadvantaged" by geolocation algorithms: our results suggest that rural users are harder to "find" in an automated fashion, preserving privacy. Geolocation inference is already employed by at least some social surveillance firms, locating tweets based on the language and metadata [**77**]. An interesting direction of research (e.g. [**125**]) is to invert the goals of this chapter and attempt to find ways reduce the "geolocatability" of a person or a population (i.e. defend against "inference attacks" [**196**]).

5.6. Limitations and Future Work

In this chapter, we necessarily binned users into categories along the urban and rural spectrum. There is without a doubt a tremendous amount of diversity in the people and behaviors present within each of these categories, and further research may want to address

this. Categorizing users based on behavior and content as opposed to demographic labels could provide additional insight into who is likely to be affected by these algorithmic biases.

Following best practices in the Twitter geolocation literature (e.g. [**138, 160, 161, 247**]), for our ground truth data, we depended on explicitly geotagged tweets for our text-based algorithm (i.e. tweet location) and a very conservative (i.e. precision-focused rather than recall-focused) geocoding of the location field for our network-based algorithm (i.e. user home location). While doing so was critical to our goal of evaluating bias in the algorithms as they were published, it is possible that these mechanisms may disproportionally remove people of one demographic relative to another demographic. Although developing a ground truth through other means (e.g. a survey) would be a major research project in its own right, examining Twitter geolocation algorithms through this lens would be a useful addition to the literature.

Another area of future work would be expanding the focus of this study (the contiguous United States) to other cultures and geographic contexts. It is known that different cultures use social media differently (and have their own population biases) so it is not clear how our results would extend to these areas. Furthermore, building on our understanding of how different populations use social media (e.g., variation of the use of mentions across cultures [**98**] will enable better prediction of where algorithmic biases might arise. Along the same lines, we sought to choose representative algorithms, but different algorithms may perform differently and introduce their own structural biases.

Finally, now that this chapter has established the role of structural bias, a very important direction of future work is finding ways to reduce or eliminate it in important algorithms. We expect that doing so could lead to an interesting and fruitful line of work.

5.7. Conclusion

This research improves our understanding of algorithmic biases in social media-based algorithms. We demonstrated the degree to which these algorithmic biases arise from both population biases in the training data and structural biases inherent to the algorithms themselves. Through the implementation of both a text-based and a network-based algorithm for geolocation inference, we found that network-based approaches may be less susceptible to structural biases. We also discussed the implications of our findings for designers and users of social-media based algorithms. These implications include (1) the need for more work developing algorithms that avoid the structural biases we observed here and (2) that global evaluation metrics can mask significant underperformance for certain populations in these algorithms.

CHAPTER 6

Transition

The first half of this dissertation focused on the representation of urban and rural communities within user-generated content in the context of three main consumers of this content: research (§3), users (§4), and algorithms (§5). Across these three studies, it was shown that *equal participation does not lead to equal representation* for these communities. Given that greater participation and more data does not guarantee equitable technologies, I turn my attention to the design of these technologies to better understand how they encode structural inequalities. The hope is that by approaching the design of these technologies from the standpoint of inequality, we might bring them closer to a place where equal participation could mean equal benefits.

Specifically, in the second half, I focus on the domain of geographic algorithms, building most directly on the findings from Chapter 5 about urban-rural bias in geolocation algorithms. The choice to focus on geographic algorithms is motivated both by 1) the large economic and social implications of geographic algorithms, and, 2) the challenges that being "geographic" brings when it comes to evaluating these algorithms to determine if they are "fair" (see §2.4). Through a number of case studies with specific geographic algorithms—vehicle routing (§7), place recommendation (§8), and geographic representation learning (§9)—I seek to provide a framework for how to evaluate this class of algorithms and some early results about how they encode structural inequalities.

CHAPTER 7

Externalities of Vehicle Routing

In this chapter, we explore how to quantify the impacts (both on drivers and communities) of vehicle routing algorithms. We use this framework to explore the impacts of changes to these algorithms given a shift of interest away from providing the fastest route between two points and towards providing paths that optimize for alternative criteria.[1]

7.1. Introduction

The simple act of driving from one place to another is an incredibly common part of many people's lives. However, it is also a surprisingly complex task: in many cases, there are seemingly countless routes between a given origin and destination pair. While the predominant focus of the literature and applications in the geographic routing domain has historically been on minimizing travel time or distance (e.g., [**20, 104**]), researchers and practitioners have recently shown interest in alternative routing criteria. For instance, researchers have developed routing systems that generate "scenic" routes (e.g. [**251, 264, 347**]), simpler routes (e.g. [**70, 279**]), and safer routes (e.g. [**92, 168, 277**]), among other alternative route optimizations (e.g. [**186, 192, 280, 352**]). Similarly, the routing platform Waze has begun to suggest routes, at least in Rio de Janeiro, that avoid areas deemed to have higher rates

[1]The work presented in this chapter was originally published in: **Johnson, I.**, Henderson, J., Perry, C., Schning, J., and Hecht, B. Beautiful...but at What Cost? An Examination of Externalities in Geographic Vehicle Routing. *Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies (PACM IMWUT / UbiComp '17).* Vol. 1, No. 2, Article 15.

of violence [**245**], and Microsoft owns a patent for pedestrian routing that avoids "unsafe" areas based on crime and weather factors [**299**].

Despite this growing interest in alternative routing strategies, there has been no controlled evaluation of the *externalities* that arise when more traditional optimization criteria such as travel time are superseded by new optimization criteria such as safety or beauty. Evaluations of these new criteria have solely considered the direct trade-off with travel time or distance, e.g. increased travel time to achieve more scenic or safer routes. These evaluations, however, miss the externalities that may arise with these new criteria, externalities that may have significant social, economic, and safety implications. For instance, at the community level, these routing approaches may lead to increased or decreased traffic in certain areas. Additionally, at the route level, these approaches may lead to routes with more turns (directly contradicting user preferences [**107, 202**], increasing driver stress [**310**] and cognitive load [**126, 236**], and potentially decreasing safety [**227, 245**]. Moreover, anecdotal evidence suggests that these externalities may be substantial. Consider, for instance, the widespread outcry about increased traffic and noise in previously out-of-the-way neighborhoods attributed to routing algorithm changes by Waze (e.g. [**143, 208, 337**]).

In this chapter, we aim to present a more nuanced and robust evaluation of routing algorithms that builds on the academic literature and burgeoning societal discussions around wayfinding and routing technologies. To do so, we developed a controlled experimental routing platform and used this platform to investigate externalities that arise with three common approaches to alternative routing: scenic routing, safety routing, and simplicity routing. Taking into account the importance of geographic context in evaluating geographic technologies (§2.5.4), we examined routes from four cities—San Francisco, New York City, London, and Manilla—and asked the following research questions:

RQ1: Does optimizing on alternative criteria in routing algorithms lead to *route-level* externalities such as more complex routes?

RQ2: Does optimizing on alternative criteria in routing algorithms lead to *community-level* externalities such as increased or decreased traffic in certain areas?

Additionally, as noted above, at least one routing platform (Waze) has already implemented alternative routing techniques [**206**], and Microsoft has a patent on similar approaches [**299**]. As such, we also saw an opportunity to use our experimental platform to better understand (and track) the criteria on which popular routing platforms are optimizing and to assess whether there are any externalities associated with these criteria. Thus, in the tradition of the algorithmic auditing literature (e.g. [**9, 41, 301**]), we also asked a third research question:

RQ3: Is there evidence of alternative routing criteria being used by *popular routing platforms*? If so, what are the associated externalities?

Overall, we find that there are large externalities associated with alternative optimization criteria and that these externalities could have substantial impacts on our communities and on the nature of the routes we use. For instance, our evidence suggests that scenic routing removes vehicles from highways (where city planners generally hope to route traffic) and redirects them to parks, popular areas, and, in some cases, wealthier areas. Scenic routing also creates substantially more complex routes involving more turns and intersections, both of which are known to make routes less desirable [**107, 189**] and are associated with negative outcomes (e.g. traffic accidents [**227**], decreased usability [**236, 310**]. Additionally, safety routing creates highly local but large impacts, redistributing traffic from pre-defined banned areas to highways and surrounding thoroughfares.

Importantly, these externalities can arise even when increases in travel time appear minimal, providing evidence that externalities may be transparent under the standard paradigm for evaluating alternative routing approaches. Along the same lines, we also identified evidence that there is substantial variation in the externalities of a given algorithm across different cities and areas within a city. As such, our findings suggest that alternative routing research must involve carefully controlled evaluations across broadly diverse geographies to fully understand the costs and benefits of an algorithmic change.

The algorithmic auditing component of our work reveals that Google Maps and MapQuest likely incorporate some non-fastest-path optimizations in their routing algorithm (e.g. they generate simpler routes that spend more time on highways). However, we found no evidence (yet) of any major externalities relative to fastest-path routing, indicating, for instance, that Google Maps and MapQuest have not implemented features like Waze has in Rio de Janeiro and applied them at scale. Our methods will allow us to easily monitor this result over time to assess if this changes (e.g. if one of these routing providers begins to route people around neighborhoods with higher crime rates).

Finally, in the spirit of work such as Jurgens et al. [**161**], which also sought to standardize the evaluation of an algorithm family (geolocation inference algorithms), we are releasing our alternative routing and evaluation platform to facilitate improved evaluations and comparisons in this domain. In addition to allowing researchers to easily consider externalities when evaluating new routing algorithms, our platform also addresses issues in the alternative routing literature such as a lack of standards, very limited open alternative routing implementations, and inconsistent evaluation criteria. We have designed our platform to be completely open-source and easily extensible (e.g., to other optimization criteria, geographic

Table 7.1. A selection of categorized alternative routing papers.

Category	Sub-categories and Papers
Positive	**Scenic** (El Ali et al. 2013, McGookin et al. 2015; Quercia, Schifanella, and Aiello 2014; Runge et al. 2016; Traunmueller et al. 2013; Zhang, Kawasaki, and Kawai 2008; Zheng et al. 2013)
Negative	**High Crime** (Elsmore et al. 2014; Fu, Lu, and Lu 2014; Kim, Cha, and Sandholm 2014; Shah et al. 2011), **Weather** (Y. Li et al. 2014), **People** (Posti et al. 2014)
Topological	**Simplicity** (Duckham and Kulik 2003; Haque, Kulik, and Klippel 2006; Shao et al. 2014), **Health** (Sharker, Karimi, and Zgibor 2012), **Efficiency** (Ganti et al. 2010)
Personalized	Letchner, Krumm, and Horvitz 2006; Delling et al. 2015; Pang et al. 1995; Ziebart et al. 2008

scales, or externalities), with the hope of supporting future research and discussion of what is important in evaluating geographic vehicle routing.

7.2. Related Work

In this section, we discuss research that motivated this work. This research emerges primarily from four areas: the large literature on alternative routing criteria, investigations into route preference, evaluation approaches in alternative routing, and the algorithms underlying commercial mapping platforms. Notably, though the research in this chapter focuses on vehicle routing, we include in our discussion of related work approaches that have considered other modes of transportation as well.

7.2.1. Routing Using Alternative Criteria

While there is a large and growing literature on developing alternatives to shortest and fastest path routing, there has not yet been an effort to summarize this literature. As such, we conducted a survey of the literature and found that the alternative routing approaches largely fall into four categories: *positive*, *negative*, *topological*, and *personalized* (see Table 7.1 below for examples of each).

The first two categories, positive and negative, are defined by the work of Golledge [**107**], which examines in part the impact of environmental features on route preferences (e.g. parks

as positive, waste dumps as negative). The third category, topological, encompasses criteria such as simplicity or driving efficiency that can be derived from basic information about the road network. The final category is personalized routing, which involves learning the personal preferences of a driver (e.g. road or turn types) and designing routes that adhere to these preferences.

Within the routing algorithms literature, positive routing often takes the form of "scenic" routing. Scenic routing has been implemented in a number of ways, e.g. reweighting edges based on an assessment of the "scenicness" of their surrounding area [**305**], adding waypoints from scenic areas near the shortest-path route [**75, 264, 342**], by generating many paths and then choosing the most scenic [**251**]. Additionally, there are also examples of optimizing for other positive criteria such as "happiness" and "quiet" [**251**] and projects peripheral to alternative routing that propose means of sensing criteria such as desirable smells [**252**] for future use in routing.

Negative routing seeks to provide routes that avoid undesirable areas. Though Golledge used waste dumps as a proxy for this type of routing, the literature largely focuses on avoiding unsafe areas as defined by high incidences of violent crime [**78, 92, 168, 277**]. Additional applications include avoiding dangerous weather [**192**] or other people [**246**]. Microsoft owns a patent [**299**] for pedestrian routing that avoids unsafe areas. Waze has already included the option to avoid high-crime areas, specifically in Rio de Janeiro, Brazil, ahead of the 2016 Olympics [**245**]. Waze also defaults to routing individuals around certain settlements that are off-limits to Israelis [**294**]. These algorithms start with the shortest path and then either add waypoints as needed to reroute away from any areas deemed undesirable or reweight edges in these areas so that they are perceived as very high cost by the algorithm.

Topological routing describes approaches that seek to optimize on some aspect of the road network itself (rather than the environment around the road network). The most common approach in the literature – other than the more traditional travel time and distance criteria – is some form of "simplicity" routing. At their core, simplicity routing approaches seek to model the ease with which a driver can follow a route, but they operationalize simplicity a number of different ways. [**107, 202, 189**] modeled simplicity as minimizing the total number of turns. Algorithmic implementations of simplicity routing have taken the approach of modeling simplicity not just using turns, but as a function of the degree of each intersection and what action is taken at that intersection (i.e. go straight or turn) [**70, 126, 279**].

Finally, personalized routing approaches generally seek to learn and model an individual's route preferences, i.e. when a user usually deviates from the fastest route and, in some cases, why s/he does so [**60**]. These models require extensive positioning (i.e. "GPS") data from drivers in order to determine these preferences. A common approach is to learn implicit preferences for specific roads [**186, 352**], though a recent approach by Delling et al. [**60**] explicitly learns the weights that each individual driver appears to give to various topological criteria (e.g., number of lanes, type of road).

We included in our experiments the most common form of alternative routing approaches in each category, with the exception of personalized routing. More specifically: for positive routing, we implemented *scenic routing*; for negative routing, we implemented *safety routing*; and for topological routing, we implemented *simplicity routing*. We did not consider personalized routing because the open "GPS" trace datasets that would be required to include personalized routing in our experiments do not exist.

7.2.2. Preferences for Alternative Criteria

Researchers have long known that fastest path and shortest path are not the only criteria on which people want to optimize their routes. Much of this knowledge emerges from a variety of surveys and field studies. For instance, in an influential paper in the field of geography, Golledge [**107**] sought to quantify the importance of various criteria in route selection. Golledge identified that minimizing distance and travel time were the most important factors, but minimizing the number of turns and maximizing the scenic/aesthetic value were also key criteria that seemed to affect which route a participant selected. Li and Wu [**189**] surveyed commuters in Florida and provided support for the findings of Golledge, but also determined that safety is an important criterion. Similarly, Manley et al. [**202**] explored which criteria best explain actual routes taken by taxi drivers in London and found that a combination of shortest distance and fewest turns was most predictive of route choice. In addition to the preference for fewer turns, increased route complexity also raises safety concerns [**227**] and has been directly tied to greater cognitive loads [**236**] and stress [**310**] for the driver. This literature, though sparse, further motivates our choice of the three alternative criteria that we consider in this work (beauty, safety, and simplicity).

7.2.3. Evaluation in Alternative Routing

As is often the case in new computing research areas (e.g. geolocation inference [**161**]), evaluation in the alternative routing literature is a highly heterogeneous process that makes comparisons between approaches difficult. Evaluations typically involve examining routing in 1-2 cities using a small number of routes, with the only evaluation metric being travel time or distance. Our work is the first to our knowledge that explicitly considers additional

evaluation criteria. In other words, this work sheds new light on the *externalities*, or side-effects, that arise with the use of alternative routing optimization criteria. Hints to the existence and importance of these externalities come from popular media, as discussed below. Like was the case in Jurgens et al. [**161**] for geolocation inference, our goal in this chapter is to conduct experiments that afford a more direct and nuanced comparison between approaches, enabling a more robust understanding of the externalities associated with each approach. We also examine routing in four cities with diverse geographic contexts, affording a broader view of how geography and algorithms interact that provides important new insight.

7.2.4. Commercial Mapping Platforms

Since MapQuest began providing online directions in 1996, most online mapping platforms have defaulted to providing the "fastest" route between a given origin and destination. The exact details of the routing algorithms being used are proprietary, often including whether or not these algorithms are optimizing on criteria other than just travel time. Some information, however, has been made public. Delling et al. [**60**] note that Microsoft Bing's algorithm takes into account dozens of topological features such as the type of road, number of lanes, speed limit, and historical traffic data, and that the algorithm optimizes for simplicity as well by incorporating turn costs. Waze became infamous for especially accident-prone turns across traffic (an externality that likely arose as a result of optimizing more heavily than other commercial routing platforms on minimizing travel time) and has also since begun to explicitly optimize for simpler turns [**245**].

The introduction of the avoidance of "dangerous" areas in Rio de Janeiro by Waze [**206**] represents a large deviation from the fastest route paradigm. Waze also has incorporated

avoidance of areas that cannot be entered legally by certain individuals, e.g. various settlements when driving in Israel [**294**]. Questions have been raised as to whether this type of safety routing merely enforces stereotypes and unfairly removes traffic (including potential retail customers) from poorer areas [**206, 245**]. These concerns were also raised when a Microsoft patent that describes a means for helping pedestrians avoid areas where crime has been reported became public [**124, 227**].

The Waze platform has also been accused of routing its users through many previously low-traffic neighborhoods. This has resulted in a number of externalities – including extensive frustration among residents (e.g. [**143, 208, 337**]) – and has led to calls for regulation of where mapping platforms can direct traffic [**208**]. Relatedly, a simulation study found that near-universal usage of fastest path routing during high-traffic times led to the redistribution traffic away from highways and onto more local roads [**300**]. Our research extends our understanding of traffic redistribution externalities to the large literature on alternative routing criteria and can help inform the policy debate about mapping platform regulation.

This research also builds on work that aims to provide some transparency to large-scale geographic systems that inform how we interact with the world, such as that by Soeller et al. [**291**], who developed a system for detecting personalization of political borders on Google Maps, and Chen et al. [**41**], who examined Uber surge pricing in San Francisco and Manhattan. This research takes a similar approach, but moves towards detecting the employment of alternative routing criteria (e.g. crime) rather than political borders or geographic biases in the sharing economy.

7.3. Methods and Framework

In order to conduct a controlled evaluation of the externalities associated with alternative routing criteria, we needed three main components: the routing algorithms for each alternative criterion (i.e. beauty, safety, simplicity), a set of origin and destination pairs, and metrics to analyze the different externalities. For each origin-destination pair, and in aggregate across all origin-destination pairs for a city, we are then able to directly compare the routes generated by each algorithm. Below, we describe in detail our implementations of each of these components.

7.3.1. Alternative Routing Algorithms

We implemented three alternative routing approaches as well as a more traditional fastest-path algorithm to provide context when necessary. For our alternative approaches, we selected scenic, safety, and simplicity routing. As noted in Section 7.2.1, these three approaches have been validated by Golledge's work and have been the subject of substantial interest in the alternative routing algorithm literature. There is no consensus in this literature, however, on *how* to operationalize criteria like "scenicness" (which is referred to as "beauty" in some literature), safety, or simplicity in a routing algorithm. Additionally, there is also a lack of open implementations of these and other alternative routing approaches.

To address these issues, we developed our own framework that consists entirely of open-source software components and publicly-available data. We have released this framework for others to use and improve[2]. As noted above, the primary goals of the framework are (1) to make it easy for researchers and developers to consider externalities in their routing algorithm evaluations and (2) to provide a greater degree of standardization in routing

[2]https://github.com/joh12041/route-externalities

algorithm evaluation more generally. Our framework is straightforward and has the benefit of being easily extensible to include additional alternative routing approaches and additional externality metrics not discussed in this chapter (e.g. number of stop lights [**107**] or the diversity of neighborhoods along the route).

For implementation of our alternative routing and fastest-path algorithms, our framework utilizes the bidirectional Dijkstra implementation provided by the open-source GraphHopper Java library[3]. GraphHopper includes standard pathfinding algorithms and imports OpenStreetMap[4] road networks to build the underlying graph for routing. GraphHopper does not take traffic into account when determining the fastest path and instead bases travel time on road speed limits included in the OpenStreetMap data. All of our alternative routing approaches are included in our released framework.

7.3.1.1. Scenic Routing. Our implementation of scenic routing is designed to replicate the approach described by Quercia et al. [**251**], except where required by the nature of our study. Broadly, Quercia et al. collect the top k shortest paths between two points and then select the path that optimizes for beauty based on data derived from Flickr photo tags. We chose this approach because it is scalable to many geographic regions and complements open-source approaches as it relies on public social media data.

Quercia et al. base their underlying data on a LIWC-based text analysis of the tags on geotagged Flickr photos. Specifically, they generate a 200m-by-200m grid across the city of interest in which each grid cell has a score based on its corresponding geotagged photos tags. They validated this approach with crowdsourced ground-truth beauty rankings. The notable variations from Quercia et al. in our system are as follows: our grid cells are slightly

[3]https://github.com/graphhopper/graphhopper
[4]https://www.openstreetmap.org

smaller and are not perfectly square (0.001 degree by 0.001 degree to necessarily speed up the algorithm) and we use Empath [**82**], a validated open-source replacement for LIWC). In order to achieve better spatial coverage, we also add geotagged tweets[5] to the Flickr [**302**] tags that are used to score each grid cell. Despite these changes, as validation we note that we see similar trade-offs in travel time to those reported by Quercia et al.

In addition to defining scenicness, the Quercia et al. approach also needs a means of generating and ranking routes based on this alternative criterion. Quercia et al. use Eppstein's algorithm, which finds the k-shortest paths between an origin and destination. We do the same through GraphHopper, but find the k fastest paths where we cap k at 1,000 as Quercia et al. demonstrated diminishing returns with larger k values (we also examined k=10,000 and found similar results to those presented below, but with greater effect sizes). As is done in Quercia et al, each route is scored as the average beauty score of the grid cells through which it passes. The route with the highest average beauty score is selected and returned.

7.3.1.2. Safety Routing. We implement safety routing as closely as possible to descriptions of the technique used by Waze in Brazil. Though the specific details of the algorithm are not public, Waze notes that it avoids areas that have "higher-than-average homicide, car robbery, or drug trafficking rates" [**206**]. We focus our efforts with respect to safety routing on New York City[6] and San Francisco[7], both of which provide public crime data. We include data from all of 2016, retaining only the crimes that overlap with the categories mentioned above. It was also noted that Waze disregarded areas that had high numbers of drivers under the assumption that their users considered these areas to be safe [**206**]. Lacking this

[5]https://dev.twitter.com/streaming/overview
[6]http://www.nyc.gov/html/nypd/html/analysis_and_planning/historical_nyc_crime_data.shtml
[7]https://data.sfgov.org/

(private) data, we implemented a proxy: we disregarded highways (speed limit greater than 70 kilometers per hour) when avoiding roads in these areas.

Waze has not released the delineations of the areas in Rio de Janeiro that were designated "unsafe," so we tested several thresholds for determining which areas to instruct our algorithm to avoid. Waze chose 25 areas in Rio de Janeiro that are described as varying in size between a block and a neighborhood [**206**]. As such, we aggregate the crime data to census tracts, which in cities are generally of a size between a few blocks and a neighborhood. We then define a threshold for the average number of crimes (normalized by the area of the census tract) such that a certain percentage of census tracts (i.e. those above the threshold) are marked as "unsafe." We tested different variations of our threshold such that it removes 25%, 15%, 10%, 5%, or 1% of census tracts[8]. We report results for the 10% threshold, finding the results for the other thresholds to be very similar.

With a list of census tracts that exceeded the threshold for crime, we conducted safety routing in GraphHopper by using fastest path routing but avoiding all road segments that pass through these census tracts and do not have a speed limit greater than 70 kilometers per hour. Thus, a fastest path that does not pass through a blocked area will be unaffected while a path that would have passed through a blocked area is rerouted to the fastest path that does not pass through a blocked area.

7.3.1.3. Simplicity Routing. We implemented simplicity routing as described by Shao et al. [**279**]. This approach scores the simplicity of a route as the sum of the complexity of each intersection through which it passes. The complexity of an intersection is modeled based on the degree of the intersection—i.e. number of intersecting roads—and what action

[8]Actually operationalizing on all areas with "higher-than-average" crime would have blocked far too many census tracts – i.e. about a third of census tracts in each city.

is to be taken by the driver—i.e. going straight or turning). We re-use the k-shortest-paths framework from scenic routing, but instead of selecting the path with the highest average beauty, our simplicity algorithm selects the path that has the lowest complexity score (i.e. the simplest route).

7.3.2. External APIs

We also included routes from two external mapping platforms[9], Google and MapQuest, to address RQ3 (evidence of alternative criteria in routes from third-party platforms). We report results for routes that were gathered on weekdays from 5-7:30pm local time for both platforms, a time of high traffic. We also gathered routes from 2-4:30am local time (weekdays) for low traffic directions, but do not report these results as we found little difference in the actual routes (i.e. ˜98% overlap in routes, just different expected travel times).

7.3.3. Origin-Destination Pairs

In order to robustly compare the routes provided by different routing optimizations, we needed a set of origin-destination coordinate pairs in each of our four cities. Ideally, analysis of routes would be done with a representative sample of route requests (e.g. from Google Maps or MapQuest server logs). However, this type of data is not publicly available. To address this issue, we take two approaches: (1) we adopt a common practice in the literature (e.g., [**126, 278, 353**]) and test the algorithms on randomly-generated origin-destination pairs from across a city's entire road network and (2) we also use publicly-available taxi trip datasets where available.

[9]Waze does not provide a public API.

With respect to our randomly-generated dataset, we generate approximately 5000 origin-destination pairs each for San Francisco (California, USA), New York City (New York, USA), London (England), and Manila (Philippines). These cities were chosen to provide regional variation while still having sufficient English speakers to provide a good source of photo tags and tweets for our scenic routing algorithm (Empath is currently limited to English).

To provide additional context when possible, we verified the validity of these randomly-selected pairs by analyzing two datasets of actual route origins and destinations based on taxi pick-ups and drop-offs, one in San Francisco [**242**] and the other in New York City[10]. As we discuss below, for our route-level externalities (RQ1), we see the same high-level findings in our taxi-sampled and randomly-generated datasets, and so we only report the results for the randomly-generated pairs. For our community-level externalities (RQ2), we reach varying conclusions depending on whether we use the taxi-sampled or randomly-generated origin-destination pairs. As such, we focus our discussion of RQ2 on San Francisco and New York City and discuss both sets of results.

7.3.4. Externality Metrics

The first set of externalities that we examine are attributes of a route that are not traditionally considered in evaluations but that have been found to be important in how people choose routes (i.e. RQ1, *route-level* externalities). First, we evaluate the complexity of the route, which we measure in several ways: number of turns [**107, 203**], number of left (or right in London) turns [**189**], and the metric that we use in simplicity routing [**70**], which takes into account the number of intersections passed through by a route and what action is taken at each intersection (i.e. turn or go straight). We also measure the beauty of all of

[10]www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

the routes [**6, 107, 189**] — another desired property of routes as determined by Golledge — doing so in the same way as described above for our algorithmic implementation of scenic routing. Finally, due to the public outcry around Waze redirecting traffic from the highways into smaller neighborhoods, we also measured the percentage of time that each route was on highways (i.e. "motorways" in GraphHopper, which are operationalized as roads with speed limits greater than 70 kilometers per hour) and the percentage of time that each route that was on slower neighborhood roads (i.e. streets below "secondary" in GraphHopper, as operationalized as roads with speed limits less than or equal to 40 kilometers per hour). For all of these metrics, we compute 99% confidence intervals by bootstrap resampling the routes 1000 times (i.e. sampling routes with replacement from the approximately 5000 generated for each algorithm and city).

The second set of externalities relates to the *community-level* impact across all the routes considered (RQ2), with much of the motivation for these externalities arising from the public discourse around Waze [**206, 208**], i.e. analysis of how traffic might be redistributed throughout a city and whether income appears to be a factor in this redistribution. We specifically focus on income because concerns have been raised that safety routing would also lead to the avoidance of poorer neighborhoods [**124, 206**]. For these externalities, we focus on the cities in which we implemented safety routing (New York City and San Francisco, both of which also have detailed census data on income available[11]). We compute how income correlates with where an alternative routing algorithm redistributes traffic (as compared to traditional fastest-path algorithms). Specifically, for road segments that saw significantly increased traffic for a given alternative routing algorithm, we calculate the weighted average of household median income based on how much additional distance of roads passed through

[11]https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

a given census tract as compared to the GraphHopper fastest algorithm. For example, if across all origin-destination pairs, there was an additional 8 km of routes in a census tract with a household median income of $60,000 and 2 km of routes in a census tract with a household median income of $50,000, then the weighted average would be $58,000 for roads that saw increased traffic. We do the same then for road segments that saw significantly less traffic and compare. We again compute 99% confidence intervals through bootstrap resampling with 1000 iterations on the origin-destination pairs.

7.3.5. Calculating Metrics for Commercial Mapping Platforms

The MapQuest and Google APIs provide the points that comprise the route that they return (i.e. latitude, longitude of enough points to accurately convey the geometry of the route). From these points, we can easily calculate the beauty of the MapQuest and Google routes through the same beauty grid-cell framework as used with GraphHopper. However, calculating the simplicity for these routes is less straightforward because the details of each intersection are not provided by Google and MapQuest. To overcome this problem, we use the map matching process developed by Newson and Krumm [**229**], which has also been implemented in GraphHopper[12]. This process converts the points into a corresponding path on the GraphHopper road network, from which simplicity can then be calculated as before. The Newsom and Krumm approach is not perfectly accurate, however, and so we enforce that the resulting matched route must be within 5% of the length of the original route in order to be considered in further analyses. Recall is generally around 85%, with the exception of Manila at 63%, which likely reflects differences in the underlying road networks of OpenStreetMap and the commercial mapping platforms.

[12]https://github.com/graphhopper/map-matching

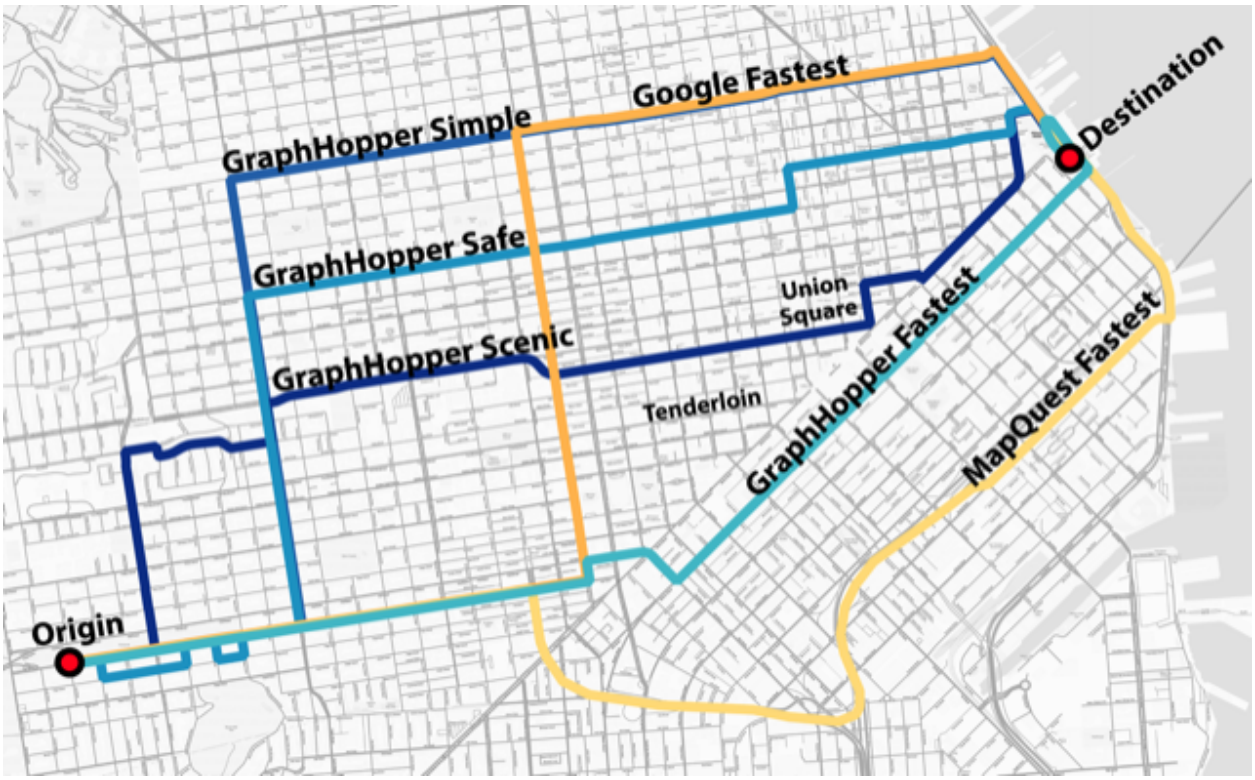


Figure 7.1. Example routes.
Routes given by each routing algorithm for an example origin-destination pair in San Francisco.

7.4. Results

In this section, we analyze and compare the routes (i.e. Google Fastest, MapQuest Fastest, GraphHopper Fastest, GraphHopper Scenic, GraphHopper Safe, GraphHopper Simple) according to the metrics described in the prior section.

7.4.1. RQ1: Route-Level Externalities for Alternative Routing Approaches

The routes corresponding to the San Francisco origin-destination pair featured in Figure 7.1 are illustrative of the type of route-level externalities that are seen between the different optimizations. In general, we see that the Google, MapQuest, and GraphHopper Simple paths

share many of the same characteristics and are longer than the GraphHopper Fastest, making more extensive use of highways. The GraphHopper Scenic route looks quite different from both the simplest and fastest routes, taking a more complicated path that passes through several popular areas such as Union Square before arriving at the destination. The GraphHopper Safe path also deviates substantially from the fastest path in order to circumvent the Tenderloin, an area of higher crime in San Francisco.

These trends in the route-level externalities that arise as a result of different optimization criteria can be seen in Figure 7.2, which shows the results of each route-level evaluation metric by Euclidean distance between the origin and destination (x-axes) and city (columns). We walk through Figure 7.2 in the sub-sections below and supplement the trends with statistics from Table 5 in [**156**], which contains the actual values for each city and algorithm when the Euclidean distance is between 10 and 11 kilometers (i.e. one slice of the data in Figure 7.2). Of note, we present the externalities both normalized to the natural baseline (e.g. GraphHopper Simple for simplicity measures) and also in absolute units when the units are readily interpretable (e.g. # of turns).

7.4.1.1. Route Complexity. Optimizing on beauty or safety substantially increases the complexity of routes, which has implications for driver safety and usability. As can be seen in rows 1 and 2 of Figure 7.2, the average number of turns and therefore complexity of a route increases significantly when comparing the fastest path to either the scenic or safer paths. Specifically, for route distances of 10-11 kilometers, the scenic path takes an additional 4-5 turns over the fastest path and the safer path takes on average an additional turn.

Similar trends hold for the number of left turns (right turns in London) on a route (not shown) and the simplicity of a route (not shown, as measured by the number of intersections and action taken at each intersection). The simplest path is then an additional 2-3 turns

shorter than the fastest path. For the safest path, this increase in complexity is the result of the added distance to circumvent a given area, but, for scenic routes, there are also significantly more steps per kilometer as well (third row of Figure 7.2).

7.4.1.2. Beauty. The scenic route is about 2-3x more beautiful (row 4 of Figure 7.2) than the other routes produced with other optimization criteria depending on the city and Euclidean distance. Notably, neither simplicity nor safety seems to correlate significantly with increased or decreased beauty.

7.4.1.3. Time on Highways and in Neighborhoods. Given concerns about the shift of vehicles away from highways and onto smaller neighborhood roads, rows 5 and 6 in Figure 7.2 consider the time spent by each route on each type of road. Scenic routes spend proportionally less time on the highway than the Google, MapQuest, or GraphHopper Fastest routes while GraphHopper Simple routes spend proportionally more time on the highway. Intuitively, this makes sense – many of the scenic spots in cities are not next to highways and taking a highway generally limits the number of intersections encountered. The magnitude of the differences varies across cities. On the low end, the scenic routes in London spend about 0.5% less of their time on the highway than the fastest path. At the high end, in San Francisco, the scenic routes spend about 9% less of the route on the highway than the fastest path, which corresponds to a 70% relative decrease in the amount of time spent on highways.

Scenic routes spend a significantly greater proportion of their travel time on slower roads, i.e. smaller roads that generally go through residential neighborhoods, foreshadowing their community-level effects highlighted below. The specific proportion of time spent on these roads varies greatly by city, but the GraphHopper Fastest, Simple, and Safe routes on average

spend similar proportions of time on these roads whereas scenic routes tend to spend 25-50% (relative) more of their travel time on these roads.

7.4.2. RQ2: Community-Level Externalities of Alternative Routing Approaches

Motivated by public concern around the redistribution of traffic and disproportionate impacts on poor (or wealthy) areas, we also examined community-level externalities that may arise from optimizing on alternative criteria. Examples of these externalities are visualized in Figures 7.3 and 7.4, which show the roads that see significantly more or less traffic in New York City and San Francisco for scenic, safety, and simplicity routing. We show both the results from the randomly-generated origin-destination pairs, which provide good coverage of the entire areas, and the taxi-sampled origin-destination pairs, which are more representative of actual route concentrations.

7.4.2.1. Distribution of Traffic. For scenic routing, areas around parks see greatly increased traffic, as do popular tourist destinations and commercial districts. As can be seen in Figure 7.4, in San Francisco, the largest increases (>100 additional routes out of the approximately 5000 analyzed in the random origin-destination pairs) occur around Golden Gate Park, the Embarcadero (popular waterfront region), Glen Park, and Mission Street as it passes through the Mission District (popular commercial district). The roads that see corresponding drops in traffic are often nearby highways or similarly large thoroughfares, which have high speed limits but are not always scenic. In New York City (Figure 7.3), scenic routing led to the largest increases in traffic (again >100 out of the approximately 5000 routes in the random origin-destination pairs) on roads that border the rivers, the roads around Central Park, and in Williamsburg (a rapidly gentrifying neighborhood in Brooklyn).

Table 7.2. Alternative routing shifts in traffic and neighborhood HMI.

Average household median income of road segments that see significantly increased or decreased traffic with each alternative optimization criteria. Using origin-destination pairs derived from taxi routes (i.e. reflective of actual travel patterns as opposed to randomly generated) often demonstrates a higher income disparity between the types of roads preferred or avoided by a given alternative optimization. City-specific differences are evident as well.

Origin-Destination Pairs	Change in Traffic	Household Median Income [99% Confidence Interval] Scenic	Safe	Simple
New York City (Random)	Increase	$56,885 [55,484-57,555]	$59,110 [59,076-59,379]	$61,881 [61,838-62,160]
	Decrease	$55,902 [55,745-56,233]	$60,338 [59,256-62,561]	$59,283 [58,992-59,961]
New York City (Taxi)	Increase	$91,737 [90,997-92,590]	$91,870 [91,485-92,505]	$77,527 [74,021-78,618]
	Decrease	$88,209 [87,377-89,607]	$98,779 [96,982-101,834]	$89,106 [87,877-91,635]
San Francisco (Random)	Increase	$93,579 [93,024-93,891]	$87,352 [86,844-87,498]	$94,203 [93,868-94,383]
	Decrease	$99,039 [98,505-100,387]	$72,566 [70,590-73,357]	$98,315 [98,002-98,858]
San Francisco (Taxi)	Increase	$97,639 [97,077-98,528]	$76,660 [74,690-77,841]	$87,688 [87,543-88,184]
	Decrease	$92,135 [91,147-94,439]	$59,279 [56,607-60,863]	$73,772 [70,242-75,430]

Again, it is largely highways and nearby thoroughfares where the greatest decreases in traffic are seen.

The changes in traffic related to safety routing are much more localized, with increased traffic in order to circumvent the blocked census tracts being redirected to highways as well as more local roads that are immediately outside of the blocked areas. In San Francisco, the region that sees the largest decrease in traffic is the Tenderloin (a poorer neighborhood very close to downtown). In New York City, the taxi-generated routes show that the bulk of the traffic redistribution would occur to avoid high-crime areas in Manhattan, though the randomly-generated pairs indicate that regions of Brooklyn and Harlem would see traffic shifted to the highways as well.

Simplicity routing leads to a large increase in the amount of traffic on highways, as they have fewer intersections, but does not appear to favor or avoid any specific areas.

7.4.2.2. Income of Neighborhoods. Given concerns about safety routing criteria avoiding poorer neighborhoods, we also computed the weighted average of the household median

income for roads that saw significantly increased or decreased traffic. The results are provided in Table 7.2.

Across the different algorithms and cities, we see mixed but persuasive results that alternative optimization can lead to large and disparate externalities in the types of areas that receive increased or decreased traffic. Scenic routing favors wealthier areas in both cities. The taxi-sampled (and arguably therefore more representative of actual traffic patterns) origin-destination pairs indicate that traffic on average moves towards wealthier regions - i.e. the average household median income of areas that see increased traffic is significantly higher than that of areas that see decreased traffic. For instance, in San Francisco for the taxi-sampled origin-destination pairs, the household median income of road segments that saw increased traffic was $97,639 while it was only $92,135 for road segments that saw decreased traffic. The randomly-sampled pairs in New York City indicate no significant correlation with traffic changes and income, though scenic routing causes traffic to move to less wealthy areas in San Francisco when using the randomly-sampled pairs.

Safety routing results in mixed effects across the two cities. In San Francisco, for both the taxi-sampled and randomly-generated origin-destination pairs, we see that safety routing moves traffic towards wealthier areas. The average household median income of areas that see increased traffic is $15,000 higher than that of the areas that see decreased traffic. In New York City, we see smaller disparities in the income of areas where traffic is redistributed, but the safety routing approach actually seems to cause traffic to shift on average towards *less* wealthy areas. Adding to the robustness of these results, we note that we see the same trends if we look at alternative metrics such as the proportion of increased and decreased traffic in areas with a household median income below a given threshold, e.g. $40,000.

7.4.3. RQ3: Alternative Criteria in External Mapping Platforms

Our results suggest that Google and MapQuest are likely incorporating simplicity as an optimization criterion in addition to travel time (like Bing [**60**]), generating simpler routes than would be expected under a pure fastest-path approach. For instance, Figure 7.2 shows that both Google and MapQuest provide routes that are similar to the GraphHopper Fastest and GraphHopper Simple routes. Interestingly, we calculated the percentage overlap between each combination of the various GraphHopper and external platform routes and found that MapQuest and Google are most similar to each other but that the highest overlap between Google or MapQuest and the GraphHopper routes is with GraphHopper Simple and not with GraphHopper Fastest.

Importantly, however, we do not see evidence of Waze-style safety routing being applied in either platform, i.e. neither Google nor MapQuest appear to be excluding any neighborhoods from their routes. More generally, we also observe no major externalities relative to GraphHopper Fastest in either commercial platform, aside from an increase in simplicity. This can be seen in Figure 7.5, which shows the significant differences (at 99% confidence) in the number of routes that pass over a given road segment when comparing Google and MapQuest Fastest versus GraphHopper Fastest. While many roads show different levels of traffic, there is no clear geographic concentration in roads that are favored or avoided. Looking back at Figures 7.3 and 7.4, we see that scenic and safety routing resulted in areas in the city where many roads all saw an increase (e.g. a popular and pretty neighborhood in scenic routing) or a decrease in traffic (e.g. an "unsafe" area in safety routing). We do not see these patterns appear for Google or MapQuest in Figure 7.5.

7.5. Discussion

In this section, we discuss the implications of the above findings for both the design of routing algorithms and for public policy.

7.5.1. Societal Impacts of Alternative Routing

A clear high-level finding in the above results is that alternative optimization criteria are associated with important externalities that have not been previously considered. For instance, our results suggest that scenic routing (and safety routing to a lesser degree) led to substantially more complex routes involving several more turns on average. Turning, and specifically turns against traffic, are known predictors of collisions [**210**] and are less preferred by users [**107**]. Additional turns also present a usability challenge, with more complex routes leading to greater cognitive load [**236**], increased driver stress [**310**], and a greater likelihood of wrong turns and increased driving distance [**126**].

We also found that, if widely deployed, alternative optimization criteria such as beauty and safety would very likely lead to some of the externalities that have been a matter of public discourse and frustration with Waze [**143, 206, 208, 337**]. Scenic routing redistributes traffic from highways into parks, popular areas, and onto slower, neighborhood roads. This raises concerns that optimizing on beauty could further contribute to the frustrations about increased traffic in previously low-trafficked neighborhoods [**208**]. These traffic increases on local roads have already led to calls by city council representatives in several cities to limit where mapping applications can direct their users [**143**]. Alongside the frustration and potential safety concerns of residents, high levels of traffic have also been tied to negative health outcomes [**235**].

By design, current safety routing approaches remove traffic from specific communities, which clearly could lead to economic and social impacts on those communities. While important to recognize, that this occurs in safety routing is not surprising. More surprising, however, is that traffic is not always just redistributed to surrounding (and likely similar) communities, but instead has far-reaching impacts and often is moved to highways and major thoroughfares that circumvent the larger area. Just as concerns have been raised about filter bubbles associated with the personalization of information consumption (e.g. [**291, 172**]), safety routing may perform a similar function, allowing people to avoid areas that they do not want to see [**206**] and potentially shaping how we perceive the world [**99**]. A more balanced approach to safety routing might implement the avoidance of these areas only at times of low traffic and, during high-traffic times of the day, actually favor the "dangerous" areas so as to reduce the potential economic and social impacts of decreased traffic and visibility of these areas. Additionally, rather than focusing on high rates of crime, safety routing might instead focus on reducing the risk of collision by avoiding more dangerous driving maneuvers or crowded areas where accidents are more likely to occur. Finally, simplicity routing might be considered as an approach that keeps drivers on highways away from neighborhoods in general without having disparate impact on just a few neighborhoods designated by a mapping platform.

In this chapter, we explored previously-proposed alternative routing criteria with the concern that these could lead to adverse and disparate impacts in specific areas. One can also imagine, though, alternative routing criteria that lead to a greater diversity of experiences for the driver and more uniform impact on neighborhoods. We invite further discussion of what other alternative criteria might be considered that would arguably lead to *positive* externalities.

7.5.2. Towards Improved Routing Evaluations

Our results also suggest that traditional routing algorithm evaluations are insufficient to capture the potential for externalities. Specifically, the sole focus of traditional routing algorithm evaluations has been on potential increases in travel-time or distance, but this can hide important negative outcomes such as increased complexity (and its associated safety effects) and undesirable traffic patterns. Additionally, focusing just on travel-time and distance can lead to the conclusion that the trade-offs of an alternative optimization diminish rapidly with distance whereas we find externalities whose effects are relatively constant across distance. As can be seen in the final row of Figure 7.2, the increased travel time costs for the alternative optimizations diminish as the route distance increases (this matches what is seen by Quercia et al. [**251**] as well). However, this drop-off in magnitude of the trade-off is not nearly as immediate or does not occur when examining certain externalities, such as the number of turns or what types of roads comprise the routes.

7.5.3. Algorithmic Auditing

In this chapter, we provided some of the first audits of routes generated by major mapping platforms. We did not find any evidence of major negative externalities associated with Google and MapQuest routes, with values for the externalities that we studied generally falling in the same range as those for GraphHopper Fastest and GraphHopper Simple.

As more commercial mapping platforms take steps like Waze has done to include notions of safety or other alternative optimizations in their algorithm, it is important that the research community continue to analyze the routes that these platforms are providing to

ensure that any negative externalities are monitored. This is especially critical given the extent to which these platforms increasingly define the movement patterns of millions of people around the world. We have built our platform to accommodate the evaluation of routes from any external source by incorporating the map-matching component that converts a series of latitude-longitude coordinates to a route on the OpenStreetMap-based road network used internally by GraphHopper. Furthermore, the dataset of routes collected in the course of this research can serve as a baseline for future evaluations in order to detect changes in the routes provided by commercial mapping platforms.

7.5.4. Geography and Algorithms

In this research, we found that the externalities associated with a given routing algorithm varied across our four cities. For route-level externalities, we saw different effect sizes in each city, but the broad trends remained consistent. Among community-level externalities though, the nature of the trends changed from city to city. For instance, safety routing redistributed traffic to less wealthy areas in New York City and to more wealthy areas in San Francisco. The dependence of externalities on local geography extended to our choice of origin-destination pairs as well — i.e. randomly-generated vs. sampled from taxi routes. For route-level externalities, the choice of origin-destination pairs did not affect the trends, but the different sets of origin-destination pairs did lead to different conclusions for our examination of community-level externalities.

These results indicate that the interaction between routing algorithms and geography, especially when evaluating community-level effects, is highly dependent on the underlying urban structure and on origin-destination patterns within that structure. Care should be taken when generalizing results from one or two cities to other settings. Especially given

the ubiquity of these algorithms (e.g. Google Maps alone has over a billion unique monthly users [**65**]), expanding alternative routing research to incorporate more geographic contexts will be important for guiding the design of these algorithms and supporting continued public discourse. We hope that our evaluation platform can assist in this endeavor.

7.6. Future Work and Limitations

One of the large questions raised by this work is how might we design alternative routing algorithms in such a way as to realize their promised benefits while reducing the associated negative externalities. The literature provides some hints that are worth exploring. For instance, the simplicity routing literature notes that hybrid, multi-criteria approaches (e.g. balancing how much weight is given to simplicity and how much is given to travel time) often greatly reduce the complexity of routes while incurring minimal time costs [**126, 279**]. Further study could examine whether multi-criteria optimizations, or other approaches that might directly consider externalities as a cost, show promise for reducing externalities. While we mentioned one way in which safety routing might be implemented in a more balanced form in Section 7.5.1, further insight might be gleaned from the social science literature (e.g. Jane Jacobs).

While the public discourse around Waze inspired some of this work, Waze currently does not provide a public API. Future work might pursue alternative means of collecting Waze routes, in part as an automatic means of detecting whether Waze routes are avoiding new areas. Furthermore, future work could also take advantage of open-source traffic data (e.g. [**216**]) to better understand the behavior of commercial mapping platform routes and explore how alternative routing algorithms react to changes in traffic as well.

We chose four cities as our geographic context for this study and sought to include cities from around the world. However, as noted in the discussion of geography and algorithms in Section 7.5.4, it is very probable that different impacts would be seen in other geographic contexts, such as in new cities or in suburban and rural areas. An interesting line of work would involve categorizing different cities based on how these algorithms perform so as to build a better understanding of how to more effectively target areas for study. In other words, are there classes of urban structures in which routing algorithms tend to have similar externalities? Of course, repeating our research in suburban and rural areas is also an important direction of future work.

We view this research as a first step towards understanding the externalities associated with various routing criteria. We sought to design our alternative routing algorithms based on actively-developed open-source libraries (e.g. GraphHopper, map matching) or published and validated methods (e.g., Empath [**82**], the k-shortest path approach developed by Quercia et al. [**251**]). However, there are many parameters and possible approaches to alternative routing, which, if executed differently, might lead to different results. Similarly, we analyzed the routes provided by Google and MapQuest for an initial directions request, but it is possible that these routes would be updated as they are driven in order to take advantage of shortcuts off of the highways. Finally, we built on previously-published methods to generate our alternative routes, but incorporating in human assessments of the resultant routes would provide additional certainty that the routes would be perceived as more scenic or simpler or safer.

7.7. Conclusion

In this chapter, we provide the first robust assessment of the externalities associated with different alternative optimization criteria in geographic vehicle routing. We show that these externalities are substantial and would not be detected by traditional routing evaluations. For instance, we find that scenic and safety routing lead to more complex routes — increasing accident risks and other negative effects — as well as substantially increased traffic in various communities. The community-level impacts vary across different cities, however, demonstrating that evaluation across multiple geographic contexts is necessary in order to understand the impact of alternative routing approaches and highlighting the complex relationship between geography and algorithms more generally. We do not find evidence of negatives externalities arising in Google and MapQuest but release our evaluation platform so as to support continued evaluation and monitoring of commercial routing platforms. Finally, we discuss how algorithm designers might better balance the benefits of alternative optimization criteria with the externalities that can arise through their use.

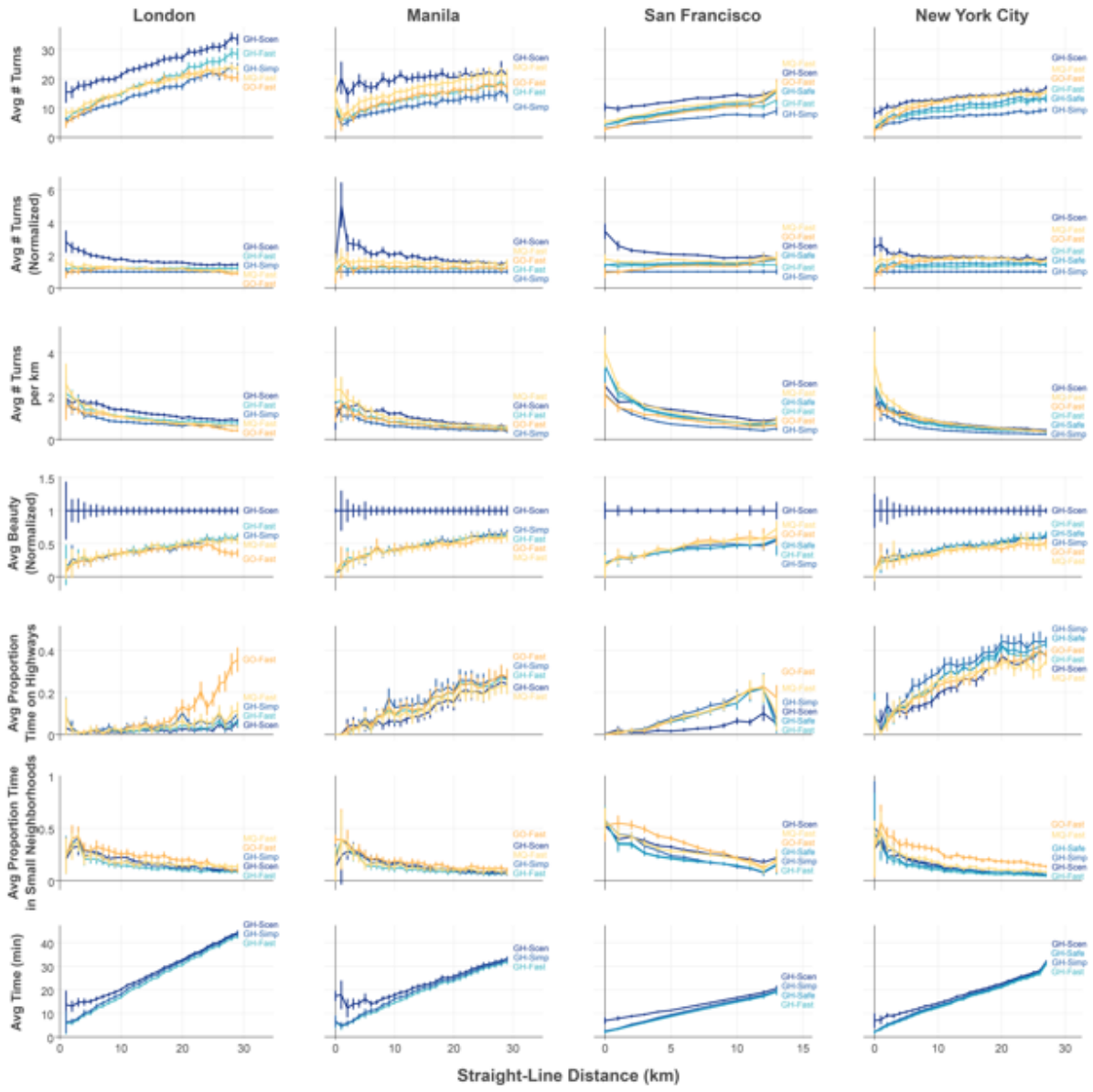


Figure 7.2. Routing externality graphs
Comparison of the route-level externalities (rows) by city (columns), distance between the origin and destination pair (x-axis), and routing algorithm (lines). 99% confidence intervals calculated through bootstrap resampling. Average travel time (bottom row) for Google Fastest and MapQuest Fastest routes are not shown because any differences between them and the GraphHopper routes likely arise from variation in the underlying data for calculating travel time

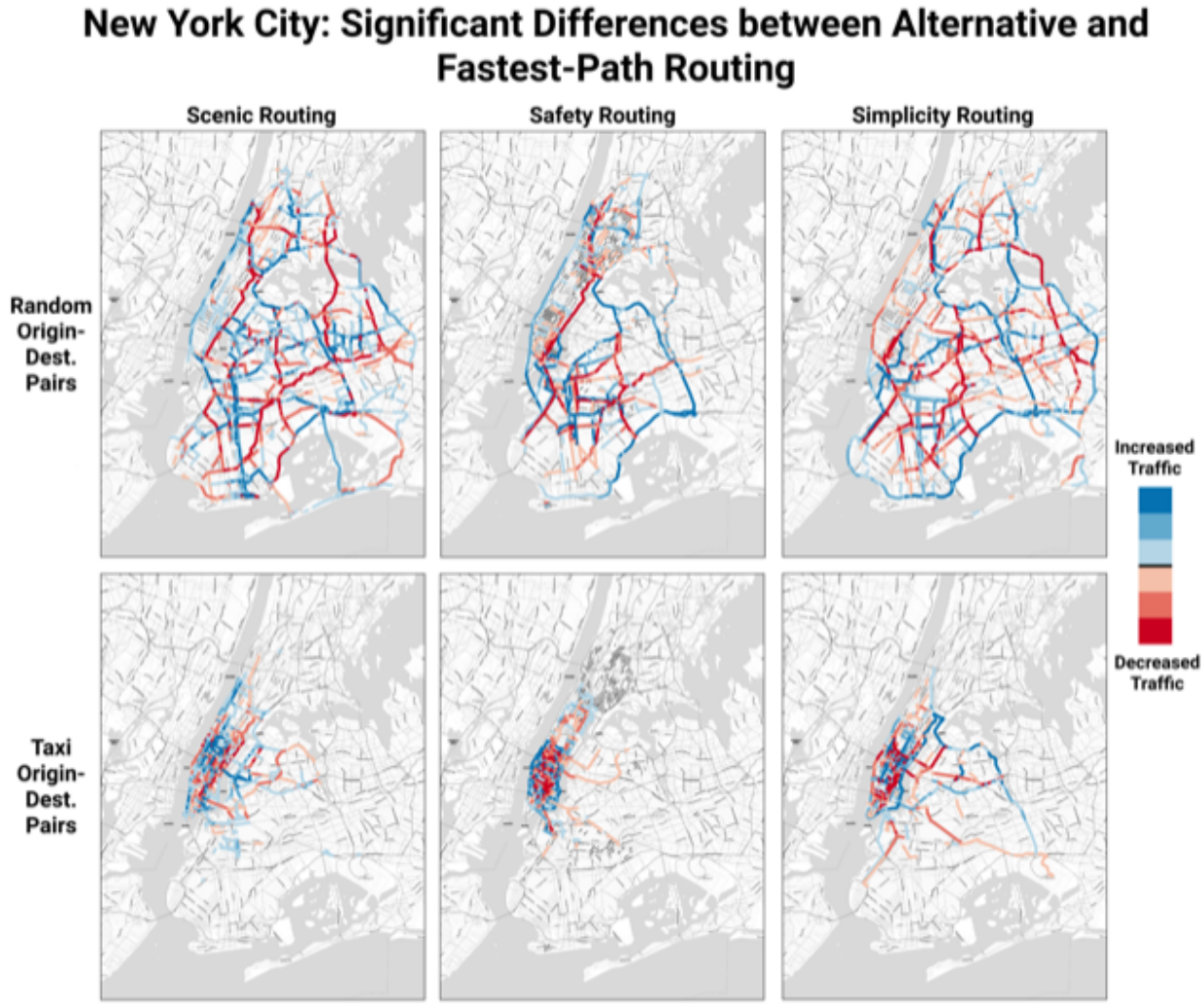


Figure 7.3. Route differences in New York City.
Comparison of alternative routing algorithms and GraphHopper Fastest algorithm, showing road segments with a significant change in the number of routes that pass over them across all origin-destination pairs for GraphHopper Scenic (left), Safe (middle), and Simple (right) routing in New York City. Results for random origin-destination pairs are shown on top and taxi-sampled origin-destinations are shown on bottom. Blue road segments had more routes pass over them with the alternative routing algorithm as compared to the GraphHopper Fastest algorithm, and red road segments had fewer routes pass over them with the alternative routing algorithm as compared to GraphHopper Fastest. The specific color thresholds were set by quantiles with the constraint mentioned above that blue represents increased traffic and red represents decreased traffic. Darker colors indicate a greater magnitude in the difference in number of routes passing over a given road segment. Black slanted lines in the GraphHopper Safe maps indicate blocked areas in safety routing.

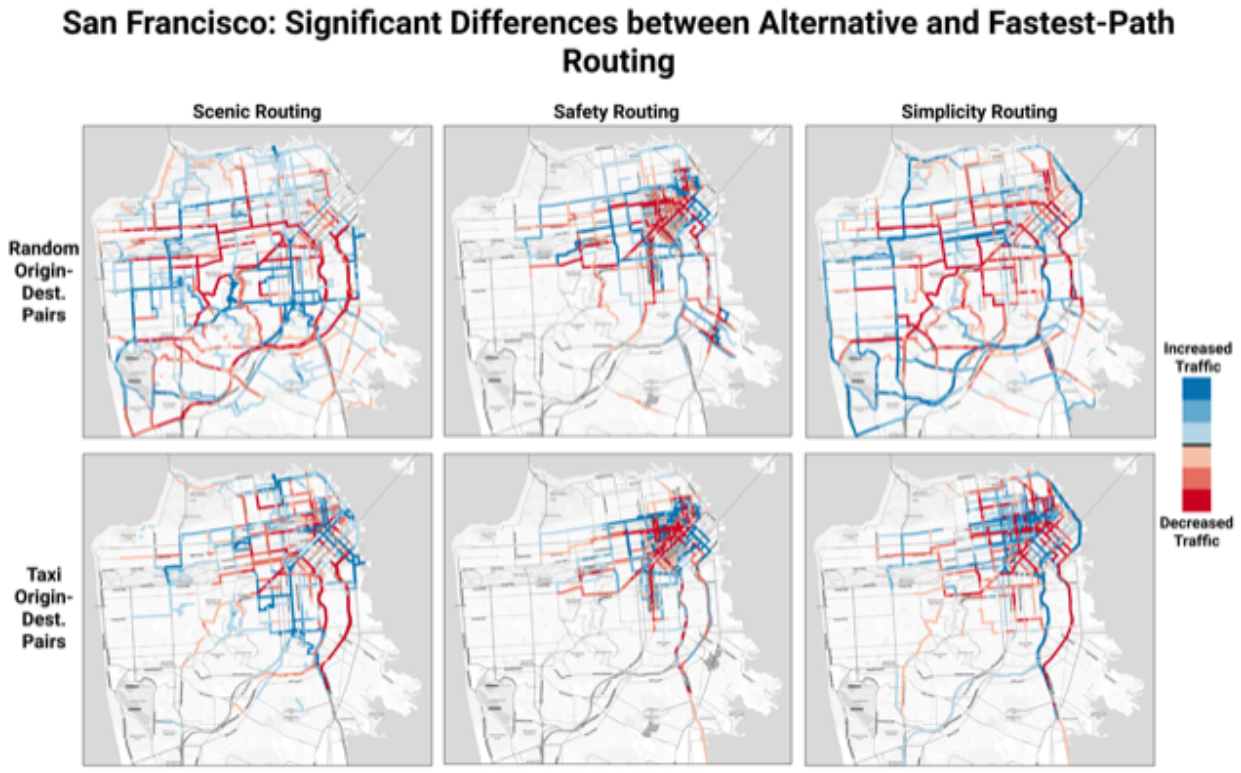


Figure 7.4. Route differences in San Francisco.
Comparison of alternative routing algorithms and GraphHopper Fastest algorithm, showing road segments with a significant change in the number of routes that pass over them across all origin-destination pairs for GraphHopper Scenic (left), Safe (middle), and Simple (right) routing in San Francisco. Results for random origin-destination pairs are shown on top and taxi-sampled origin-destinations are shown on bottom. Blue road segments had more routes pass over them with the alternative routing algorithm as compared to the GraphHopper Fastest algorithm, and red road segments had fewer routes pass over them with the alternative routing algorithm as compared to GraphHopper Fastest. The specific color thresholds were set by quantiles with the constraint mentioned above that blue represents increased traffic and red represents decreased traffic. Darker colors indicate a greater magnitude in the difference in number of routes passing over a given road segment. Black slanted lines in the GraphHopper Safe maps indicate blocked areas in safety routing.

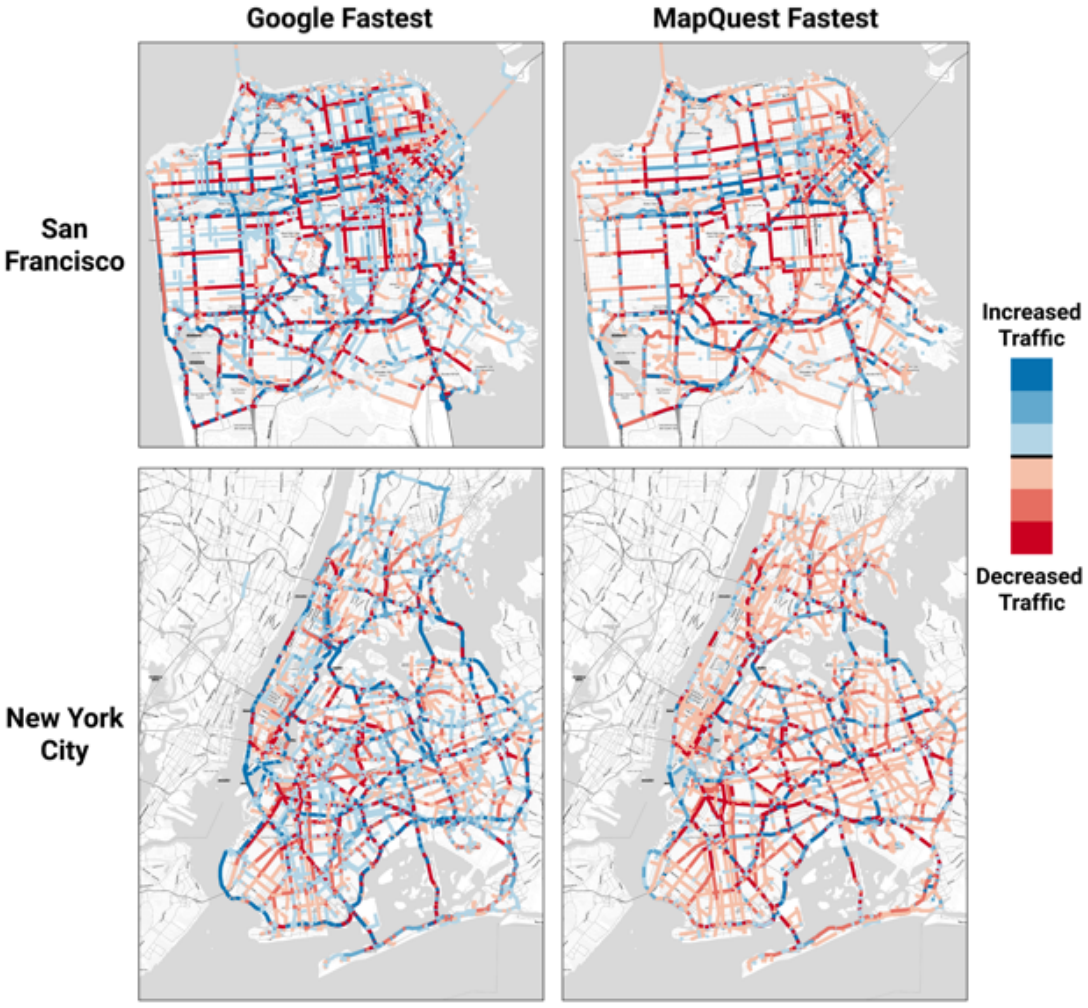


Figure 7.5. Audit of Google and Mapquest routes.
Comparison of third-party mapping platforms and GraphHopper Fastest algorithm, showing road segments with a significant change in the number of routes that pass over them across all origin-destination pairs for Google (left) and MapQuest (right) routing in San Francisco (top) and New York City (bottom). Results are shown for random origin-destination pairs. Blue road segments had more routes pass over them with the alternative routing algorithm as compared to the GraphHopper Fastest algorithm, and red road segments had fewer routes pass over them with the alternative routing algorithm as compared to GraphHopper Fastest. The specific color thresholds were set by quantiles with the constraint mentioned above that blue represents increased traffic and red represents decreased traffic. Darker colors indicate a greater magnitude in the difference in number of routes passing over a given road segment.

CHAPTER 8

Place Recommendation and Locality Bias

In this chapter, we explore the degree to which geographic collaborative-filtering models (commonly used in place recommendation) encode location. We develop methods for evaluating the biases embedded within these models and test methods for reducing this locality bias.[1]

8.1. Introduction

Place recommendation—i.e. providing users with (personalized) recommendations to visit places such as restaurants—is an increasingly powerful and commonplace technology. Foursquare[2], Yelp[3] and Google[4] all have large user bases and provide or have announced that they will be providing personalized restaurant recommendations. Researchers have studied variants of place recommendation extensively, such as predicting the rating a user will give a restaurant [**134**], next point-of-interest (POI) visited by a user [**233, 84, 331, 326, 336, 343, 313, 199, 133, 269, 329**] or the next visitors for a given point-of-interest [**83, 345, 335**].

Collaborative-filtering-based models are a common and straightforward approach to personalized place recommendations (e.g., [**331, 197, 83, 269, 334**]). At their simplest, these

[1]The work presented in this chapter builds on research conducted while I was an intern under the mentorship of Chris Welty.
[2]https://www.eater.com/2014/7/23/6182431/foursquare-reveals-overhaul-of-its-recommendation-app
[3]https://www.digitaltrends.com/social-media/yelp-collections-announced/
[4]http://www.travelandleisure.com/travel-tips/mobile-apps/google-maps-percentage-match

models represent users based on which POIs they visit and POIs by who visited them. Recommendations can then be made for a given user based on what POIs are visited by similar users. Collaborative filtering is also a natural approach to personalized recommendation in that it leverages a user's past history of POI visits to predict their future visits.

There is anecdotal evidence, however, that user-visit collaborative filtering (CF) strongly encodes location. For instance, the mobility of most people resembles Lévy-flight behavior [**32, 45**], where the vast majority of their visits occur within a relatively small region. Places that are nearby are much more likely to have overlapping users (and therefore similar representations) than places that are further away, regardless of how similar they are. Many place recommendation approaches explicitly encode geographic location to take advantage of this (e.g., [**331, 197, 83**]). Furthermore, as we show, the common approach in the literature to evaluating place recommendation models—calculating the predicted rank of a held-out visit against all or a random subset of POIs—strongly favors models that encode location.

While collaborative-filtering models that strongly encode location seemingly have higher precision at recommendation, we raise a number of concerns about this "locality bias". It is not clear to what degree a place recommendation model should encode location and we could refer to this aspect merely as "spatial representation", but we refer to this as "bias", which carries connotations of deviation from a standard level, for a number of reasons: 1) as we show in this work, models with higher locality bias do not in fact perform better at predicting which restaurant an individual attended within a given neighborhood, 2) this locality bias inhibits model generalization and means that collaborative-filtering-based models do not make meaningful recommendations in regions outside of where a given user spends their time distant regions (see §8.2.1), and, 3) given strong patterns of residential segregation within countries like the United States [**2, 286**], there is reason to be concerned that models

that strongly encode location will also reinforce these patterns. Together, these three factors suggest that the strong encoding of location in collaborative-filtering-based place recommendation models should not be treated as a natural component of these models, but instead a more spurious relationship between people and locations that these models reflect due to a reliance on user-visit trace data.

In this paper, we use public data on restaurant visits from Yelp to explore the degree to which locality bias is in encoded by CF-based methods for place recommendation. We find strong locality bias in user-visit CF. To overcome this bias, we exploit the intuition that entities that appear in multiple places but that are semantically similar can serve as bridges across this locality bias: 1) *users who travel* more extensively, and 2) *chain restaurants* that appear in multiple locations. To evaluate our findings, we propose a new method of evaluating place recommendation models that does not reward models that encode locality bias and examine the relationship between locality bias and fairness.

Specifically, we ask the following research questions:

- **RQ1**: How strongly do user-visit collaborative-filtering embeddings encode location over other concepts such as a restaurant's category?
- **RQ2**: How might we reduce this locality bias in user-visit collaborative-filtering embeddings?
- **RQ3:** What is the relationship between locality bias, recommendation accuracy, and recommendation fairness?

We make the following contributions:

- **Locality Bias**: We provide a robust quantitative evaluation of the degree to which collaborative-filtering-based techniques encode physical location as opposed to content-based metrics such as restaurant category.
- **Place and User Bridges**: We demonstrate two approaches for reducing this locality bias within a region without substantially decreasing recommendation accuracy.
- **Place Recommendation Evaluation**: We introduce a new method for evaluating place recommendation models that does not favor models that strongly encode locality bias.
- **Provider Fairness**: We show that reducing locality bias in the embeddings and each user's recommendations does not lead to increased provider fairness—i.e. a more equal spatial distribution of recommended restaurants across all recommendations.

8.2. Related Work

This work builds on extensive literature learning feature embeddings from co-occurrence data and how these methods can encode bias that is present in the underlying data. This bias is explored in the context of place recommendation and provider fairness (see §2.4.2).

8.2.1. Place Recommendation

Personalized place recommendation (also known as POI recommendation, next place prediction, and location-based recommendation) has attracted substantial interest from the research community.[5] At its simplest, this task requires representations of users and representations of POIs such that a model can recommend a likely POI for a given user based

[5]See [**333**] for a survey of methods.

upon what is known about them. A common approach to this task is collaborative filtering over user-visit data (e.g., [**331, 334, 133, 42**]), where users are represented based on where they have visited and restaurants are represented based on who has visited them. Intuitively, the model recommends a POI to a user if similar users have visited that POI. These models often incorporate additional structural variables such as time-of-day (e.g., [**334, 269**]) or inferred categorical preferences (e.g., [**343, 133**]) and alternative models for performing dimensionality reduction (e.g., Bayesian methods [**257**], metric embeddings [**84**]), but at their core they all model users based on their past visit behavior and users who visit the same POIs are represented similarly.

A major challenge for these recommendation models is whether they simply learn the areas that a user frequents or can effectively generalize what is inferred about a user's tastes in their main area of activity to more distant areas. This question is important to understanding how well these models generalize and adapt to users who move or wish to explore outside of their normal patterns. Models that primarily learn location would also be in danger of reinforcing existing patterns of segregation. User mobility data—i.e. the locations of the POIs visited by a particular user—displays strong spatial regularity, with most visits happening within a small physical region [**233, 331**] and only occasional long-distance deviations, as modeled by Lévy-flight behavior [**32, 45**]. Many POI recommendation models explicitly leverage this locality bias, or the fact that the majority of a user's visits happen in a small region, to favor nearby POIs (e.g., [**331, 84, 336**]).

The most related body of research to the question of whether geographic recommendation models are truly learning user tastes or largely memorizing location is that of multi-region recommendation. Research that has explored these questions has largely moved away from user-visit CF models and depended on topic modeling (e.g., cuisines or other tags associated

with a POI) or non-personalized attributes such as time-of-day. Yin et al. [**332**] introduce their location-content-aware recommender system (LCARS), which builds a topic model for each POI and models a user's preference for these topics, allowing them to make predictions in new cities for a user. Zhang and Wang [**341**], Xu et al. [**327**], and Wang et al. [**315**] also model places based upon tags associated with them and use this information to make predictions in distant regions for which traditional user-visit CF methods would fail. Maeda et al. [**199**] make predictions in unfamiliar areas (to the user) based upon a place's overall popularity, distance to the user's previous location, and contextual factors such as weather, but do not take a user-visit CF approach that could leverage a user's past history of visits. Ference et al. [**85**] build a user-visit CF model but also incorporate a user's social network in determining similar users for making more distant recommendations.

This body of research is unsatisfying for two reasons. First, it approaches this challenge primarily from the standpoint of recommendation accuracy. Thus, solutions are judged effective if they accurately predict check-ins when a user travels but are not evaluated for how strongly they encode existing spatial patterns. Second, the shift to content-based topic modeling means that little attention is paid to the root problem, that of the strong locality bias encoded within user-visit data (and therefore presumably any model trained primarily on this data). This is a problem that would also likely pertain to other types of models built on user data with a strong spatial component, such as embeddings trained from local search queries or page view history.

8.2.2. Embeddings and Bias

Collaborative-filtering approaches to place recommendation generally rely on some form of dimensionality reduction (e.g., matrix factorization [**197**], word2vec-style model [**83**]) such

Table 8.1. Aggregate statistics for users and restaurants from Yelp dataset included in the analysis.

Aggregate statistics for users and restaurants from Yelp dataset included in the analysis.

Region	# of Reviews	# of Users	# of POIs	# of Chains (# Locations)
Phoenix, AZ, USA	692,124	70,724	12,657	948 (5,897)
Charlotte, NC, USA	136,543	11,941	4,323	361 (1,729)

that each place and user is represented by a low-dimensional vector—i.e. embedding. While these vector representations allow for efficient predictions, they also provide a way of examining what a model has learned about a given user or POI. For instance, in language modeling, word embeddings can encode valuable semantic relationships such as between "cartography" and "maps". However, these techniques also preserve undesirable associations that exist in the data, such as gender- [**29**] or age-related [**63**] stereotypes. Researchers have explored methods of imposing constraints on what is encoded in these word embeddings (e.g., [**29, 248**]). This affordance of embeddings has largely been unexplored in place recommendation though.

8.3. Data and Methods

The goal of this work is to explore how different choices for representing data about users and restaurants lead to embeddings that encode different types of relationships and bias. Thus, for all of the embeddings in this work, the underlying co-occurrence data is represented as a two-dimensional matrix and we hold the factorization algorithm (and hyper-parameters) constant across all experiments.

Table 8.2. Co-occurrence statistics for Phoenix, AZ.

Aggregate statistics for different approaches to representing restaurants and users via co-occurrence statistics for Phoenix, AZ. For instance, from the first row, we see that there are 67,977 users with an average of 10.1 reviews and median of 5 reviews, where reviews are taken as evidence of a visit.

Co-occurrence	Row	Vocab Size	Avg (Med) Row Sum	Column	Vocab Size	Avg (Med) Col Sum
User-Visit CF	Users	67977	10.1 (5)	Rest.	12657	54.7 (22)
Restaurant-Review	Unigrams	32793	449.8 (69)	Rest.	12657	1165.4 (836)
Restaurant-Attribute	Attributes	164	2357.1 (578)	Rest.	12657	30.5 (36)

8.3.1. Yelp Open Dataset

All embeddings are learned from data from the Yelp Open Dataset[6], a public dataset provided by Yelp that contains reviews for restaurants across 11 metropolitan areas as well as aggregate data about the Yelp venues and users who provided reviews in the dataset. We focus on Yelp reviews for restaurants[7] in the Phoenix, AZ, USA, and Charlotte, NC, USA, metropolitan areas. Hereafter, we use "Phoenix" and "Charlotte" to refer to these metropolitan areas, which also include suburbs such as Scottsdale, AZ, and Concord, NC. Only users and restaurants with at least five reviews are included. Aggregate statistics are provided in Table 8.1.

8.3.2. Collaborative Filtering Approaches

For this study, we are primarily interested in learning embeddings for each restaurant, so as to measure locality bias and explore the ability of the embeddings to capture similarities across regions. We also need to build representations of each user in order to evaluate the quality of the restaurant embeddings for personalized place recommendation. For each research question, we compare three main sources of co-occurrence statistics from which

[6]https://www.yelp.com/dataset

[7]Restaurants were defined as all Yelp venues with at least one of "Restaurants", "Bars", "Coffee", or "Food" in their list of Yelp-defined categories.

we learn these embeddings, as described below and with descriptive statistics for Phoenix provided in Table 8.2.

8.3.2.1. User-Visit Collaborative Filtering. For standard user-visit CF, each row corresponds to a unique user in the dataset and each column to a unique restaurant. A cell has a value of one if that user visited that restaurant, otherwise zero. In this straightforward CF-based approach, users are similar if they go to similar restaurants, and restaurants are similar if they are visited by similar users. This represents restaurants not by their content (i.e. what they are), but by who attends them, and therefore tends to capture the relatedness between restaurants but not necessarily similarities. For example, two restaurants that serve very similar food and have similar ambiance could share very few visitors and therefore be represented very differently.

8.3.2.2. Restaurant-Review Content-Based. In this content-based approach, each row corresponds to a unigram (e.g., "tasty", "lettuce") that can be found in the text of the restaurant reviews generated by users, and each column corresponds to a unique restaurant (as with user-visit CF). The cells contain the counts of how many times that unigram appeared in reviews about that restaurant. The reviews were lower-cased and split into tokens based on whitespace and punctuation. Stop words were removed from the reviews as well as the name of the restaurant for that review (e.g., the unigrams "burger" and "king" are removed for reviews left at "Burger King"). Only words that appeared at least 20 times across the reviews were retained. This is a content-based approach, where each restaurant is represented by what has been said about them, so restaurants with similar menus and ambiance likely will be represented similarly assuming that users discuss these aspects in their reviews.

8.3.2.3. Restaurant-Attribute Content-Based. In this content-based approach, each row corresponds to an attribute (e.g., "Has-Wifi:True", "Good-for-Groups") that Yelp has

identified for that restaurant, and each column again corresponds to a unique restaurant. The cells have a value of one if that restaurant was identified as having that attribute by Yelp, otherwise zero. This is a more high-level content-based approach to representing the restaurants given that the vocabulary is restricted to a small set of concepts identified by Yelp.

8.3.3. A Framework for Measuring Bias

There are three clear stages at which bias could be measured in place recommendation: the underlying data, the representations of restaurants and users learned by the model, and the recommendations made by the model. For each stage, we design methods for quantifying the degree to which location or restaurant-specific qualities are being encoded. Then, for a given change to a geographic hyperparameter (discussed in the next section), the effect can be measured in terms of how it changes what is or is not learned by the model.

8.3.3.1. Underlying Data. To measure how strongly the underlying data encodes physical location and restaurant-specific attributes, we compute several descriptive statistics: 1) average likelihood that any two of a users visits are in the same neighborhood, 2) average distance between sequential visits by a user, 3) average overlap in restaurant categories between any two of a users visits. These statistics serve as a baseline against which to check whether a given place recommendation approach reduces, simply encodes, or amplifies bias.

8.3.3.2. Model Embeddings. Locality or content bias is straightforward to measure in the underlying human mobility data as described above. It is more ambiguous, however, as to how to measure these biases in a vector embedding, so we take several, complementary approaches:

The first set of metrics directly measure correlations between the embeddings and attributes of the restaurants, and are therefore useful for understanding how strongly the embeddings encode patterns in the underlying data. Specifically, we measure the average correlation between the cosine similarity of two restaurant embeddings and 1) distance between the two restaurants, 2) overlap of neighborhoods[8] of the restaurants, and, 3) overlap between restaurant categories.

To provide insight into the relative prominence of restaurant location vs. attributes in the embeddings, we design a fourth metric that leverages chain restaurants—i.e. restaurants with the same name (and therefore generally ownership, menu, atmosphere, etc.) but with multiple locations. Given that two locations of the same chain restaurant should be very similar with the only difference being physical location, we can compute the average cosine similarity of chain locations and compare this to the average cosine similarity between a chain location and other restaurants in its neighborhood (locality bias) and other restaurants of the same category (restaurant-attribute bias). If the highest cosine similarity is between a chain and its other locations, that indicates than an embedding is primarily capturing categorical concepts like food type or ambiance. If the highest cosine similarity is between a chain and other restaurants in its neighborhood, however, that indicates high locality bias as location is embedded more strongly than type of restaurant.

8.3.3.3. Model Recommendations. Finally, we can also look for bias in the outputs of the algorithm. The standard evaluation of place recommendation algorithms within the research literature is designed to value models that can readily learn a users location habits. We design an alternative evaluation that effectively removes this signal and focuses on the

[8]Neighborhoods are not consistently annotated in the Yelp dataset, so we augment it with neighborhood boundaries from https://github.com/blackmad/neighborhoods

restaurant-specific qualities. For instance, if a place recommendation model is trained and evaluated on data from a given city, there are K restaurants (typically several thousand) that a user could potentially visit. If each user has n_i visits, the model might be trained on each users first n_{i-1} visits and evaluated on how well it predicts the nth visit. To do so, for each user,[9] a ranked list is generated based on the likelihood of visiting each of the K restaurants according to the model. Precision-@k is then reported, where if $k = 10$, this would be the proportion of the time where the held-out nth restaurant is in the first ten spots on the ranked list. By comparing the models predictions for all restaurants despite the fact that most people visit restaurants in only a few areas that are close to them, this evaluation approach requires models to effectively learn which areas a user visits.

We design an evaluation approach that does not reward models for learning physical location (as an extension to [**332**]). Though evaluation metrics are not a geographic hyper-parameter in the actual model (i.e. they do not figure into supervised learning), they do dictate what sorts of approaches researchers take to improve place recommendation technologies in general. In practice (e.g., Google Maps, Yelp, TripAdvisor), place recommendation models generally know a users current (or desired) location and provide a ranked list of nearby restaurants. This approach highlights how well a model learns a users specific tastes in restaurants, not their preference in location. To recreate this approach in offline evaluations, this work adjusts the evaluation to only ask the model to rank the k (we use 15) restaurants that include the restaurant visited by the user and $k - 1$ physically closest restaurants.

[9]For the embeddings that do not have users as a component (i.e. restaurant-review and restaurant-attribute embeddings), we inferred an embedding for the user based upon the average embedding from the restaurants that they had visited. We also experimented with a weighted average of their visits with more recent visits receiving more weight but saw little difference in the results.

By comparing a models precision in the classic evaluation versus the alternative, location-agnostic evaluation approach, we can see how changes to geographic hyperparameters might affect how well the model learns location vs. aspects of the restaurant that are not connected to location. Specifically, if a model performs worse on the classic evaluation but equally well or better in the alternative evaluation, this indicates that locality bias is not pure signal but also contains noise that does not improve recommendations while still reinforcing existing mobility patterns.

8.3.4. Bias vs. Accuracy and Fairness

Finally, in RQ3, we examine the relationship between bias in the model, as measured in the ways described above, and the accuracy and fairness of recommendations. Accuracy is straightforward: we simply compare the precision-@k for each model in the alternative, location-agnostic evaluation. Provider fairness (P-fairness, see §2.4.2) is a measure of how equitable is the distribution of recommendations—i.e. does the model consistently recommend the same restaurants or a more diverse selection?

Unlike traditional metrics for item diversity, which focus on the diversity of a given ranking for a user, P-fairness is an evaluation of the diversity of all recommendations over a representative set of users. Lacking a dataset of local search queries, we take our test set of Yelp users as a representative set of queries. We look at P-fairness in two ways: the entropy of restaurants recommended and entropy of recommendations in each neighborhood. Higher entropy values indicate a more uniform distribution of restaurants / neighborhoods recommended and therefore greater P-fairness. We calculate these metrics based on the rankings produced by the traditional place recommendation model. We use the traditional

place recommendation model so as to understand whether the model is favoring particular restaurants or neighborhoods when location does not constrain the recommendations.

8.3.5. Geographic Hyperparameters

This section provides a foundation for thinking about geographic hyperparameters (see §2.5 for a more general overview) in place recommendation. For our first research question, we leave these geographic hyperparameters at their default settings and measure the resulting bias in the embeddings and recommendations. For our second research question, we build on two intuitions described below, and corresponding interventions, into how to reduce the magnitude of locality bias in the user-visit CF embeddings. We simply re-run the bias evaluation tasks from RQ1 on the bias-adjusted embeddings and compare. Because both sets of embeddings have the same training set of users and restaurants for which embeddings are learned as well as held-out test set of visits for recommendation evaluation, direct comparisons can be made.

8.3.5.1. Distance Decay. At first glance, the simplest approach to place recommendation is rather geography-agnostic. Just as might be done with movie recommendation via user-item collaborative filtering, matrix factorization (e.g., Funks SVD) is used to learn embeddings of restaurants and users based on a user-restaurant co-occurrence matrix and then the dot product of a given user embedding and restaurant embedding indicates the relative likelihood that that user would visit that restaurant. This description hides some implicit geographic choices though that were made, specifically distance decay both in the co-occurrence statistics and evaluation.

In movie recommendation, it is presumably just as easy to watch one movie as any other, so the overall frequency with which all users watch a movie is a good proxy for how much

to weight the surprise when an individual user does or does not watch it. Within place recommendation and user-visit data, this assumption is more tenuous. If the data covers even a somewhat large geographic region—e.g., more than 10 kilometers across [**233**]—the distance between a restaurant and where a user generally goes in the course of their day will lead to user-specific costs for each restaurant that may not correlate with the overall frequency with which users attend a given restaurant. This imbalance between global frequency and user-specific frequency leads to the model more heavily weighting visits to restaurants close to a user even if they are quite convenient (and would be much more surprising if they had not happened) while likely underweighting more distant visits made by the user that might indicate a specific interest.

8.3.5.2. Reducing Locality Bias through Users. Our first approach to reducing locality bias in place recommendation manipulates the underlying data at the user-level. We amplify data from users who do travel, building on the intuition that visits that are further from a user's dominant locale indicate greater effort, and thus should be weighted more than a nearby visit. For each user, we compute their "home location" as the geometric medoid of their visits. Rather than representing a user's visit to a restaurant as a simple one in the co-occurrence matrix, we represent it as one plus whichever of these two distances is smaller: the distance between that restaurant and the user's home location, or, the distance between that restaurant and the user's previous visit. We refer to this approach as *distance-weighted CF*. This should reduce locality bias by more strongly connecting distant restaurants.

8.3.5.3. Reducing Locality Bias through Places. Our second approach to reducing locality bias in place recommendation manipulates the underlying data at the restaurant-level. We note that certain restaurants have more than one location—i.e. franchise chain restaurants, such as McDonald's or Ruth's Chris Steak House, as well as local restaurants

that have a few locations. The different locations of a restaurant receive a different set of users, but we hypothesize that that is mainly a function of convenience and not reflective of strong differences in food, ambiance, or other qualities between the locations. Therefore, we define all restaurant names that are associated with at least two locations as chain restaurants.[10] When a user visits a chain restaurant, we randomly choose one of the locations for that chain restaurant to record as the visit. We refer to this as *chain-swapped CF*. This should reduce locality bias by geographically spreading out a user's visits without changing their "expressed tastes."

8.3.6. Representation Learning Algorithm

To learn the CF embeddings for the different representations of users and restaurants, we take a common approach of matrix factorization over co-occurrence statistics (e.g., [**43, 327, 197**]). Specifically, to learn user and restaurant embeddings via matrix factorization, we use the Bayesian Personalized Ranking (BPR) loss [**257**] and as implemented in [**176**]. To verify the trends we saw with BPR, we also tested singular-value decomposition (SVD) and Non-Negative Matrix Factorization (NMF), both of which we found produced qualitatively similar results and identical conclusions. Both SVD and NMF have been validated for place recommendation (e.g., [**197, 333**]).

8.4. Results

In RQ1, we examine bias within the standard user-visit CF approach to place recommendation (examining the underlying data, model embeddings, and recommendation accuracy

[10]We examined this list and verified that this simple heuristic indeed largely captures restaurants that are directly related to each other, both classic chain restaurants like McDonald's and local restaurants with multiple locations.

as detailed in §8.3.3). In RQ2, we explore how the user bridges and place bridges reduce this locality bias. Finally, in RQ3, we explore how locality bias then relates to recommendation accuracy and fairness. For all estimates, 99% confidence intervals are bootstrapped via resampling with 1000 iterations. Chain restaurants are excluded from the analysis of locality bias and recommendation accuracy because they were directly manipulated in the chain-swapped embeddings. We focus on Phoenix, AZ, and Charlotte, NC, here, but found similar results across the other metropolitan areas that we tested: Las Vegas, NV, USA; Cleveland, OH, USA; Pittsburgh, PA, USA; Toronto, ON, Canada.

8.4.1. RQ1: Baseline Underlying Data

As baseline data against which to measure the bias in embeddings, we calculated location bias in the underlying Yelp data (Table 8.3). This baseline data includes both lower limit to location and category overlap, which is computed as average relationship between two restaurants that are selected at random. It also includes the likelihood of overlap as computed from the Yelp check-in training data, which is the average of the overlap computed individually for each user. We note, as expected, that users demonstrate strong spatial and category homophily, consistently choosing restaurants that are in a small geographic area and of similar type. If the bias in the embeddings fall between these two estimates, the model has decreased the bias in the underlying data. If the bias in the embeddings is approximately the same as in the underlying data, the model has simply encoded the what exists in the data. And if the bias in the model is greater than in the underlying data, it has amplified the bias.

Table 8.3. Baseline bias in Yelp check-in data

Baseline data for location and category overlap between restaurants. Each measure has low-end estimates ("Random": selecting two restaurants at random and computing the metric) and likelihoods based on the user-visit data ("User Average": average of the metric computed for each user). Neighborhood overlap is the proportion of time two restaurants are in the same neighborhood. Physical Distance is the average distance in kilometers between two check-ins. Category overlap is the proportion of the time two restaurants are in the same main category per Yelp.

City	Neighborhood Overlap		Physical Distance (km)		Category Overlap	
	Random	User Average	Random	User Average	Random	User Average
Charlotte, NC	0.062	0.438	21.423 km	10.066 km	0.073	0.357
Phoenix, AZ	0.070	0.588	25.865 km	12.963 km	0.079	0.395

8.4.2. RQ1: Locality Bias in Model Embeddings

We find consistent evidence of locality bias within user-visit CF-based embeddings as compared to our content-based embeddings. Figure 8.1 shows the proportion of time for which a given restaurant and its k-th nearest neighbor, based on cosine similarity of embeddings, are in the same neighborhood for the data from Phoenix, AZ. If the embeddings did not encode location at all, the likelihood that two restaurants are in the same neighborhood was determined empirically to be 7% in Phoenix (Table 8.3, Neighborhood Overlap - Random). For the Restaurant-Review and Restaurant-Attribute embeddings, the likelihood that a restaurant's nearest neighbor is in the same neighborhood is slightly above this baseline at 14% and 11% (Pearson correlations ρ of 0.046 and 0.014, which are significant at 99% confidence but also reflect minimal locality bias). A restaurant's nearest neighbor in the standard user-visit CF embeddings is in the same neighborhood 50% of the time though (ρ=0.268), substantially higher than baseline, but also still lower than what we see in the underlying data. This suggests that the model did not fully encode the locality bias in check-in data. We see supporting trends when examining the physical distance separating a restaurant and each of its k-nearest neighbors.

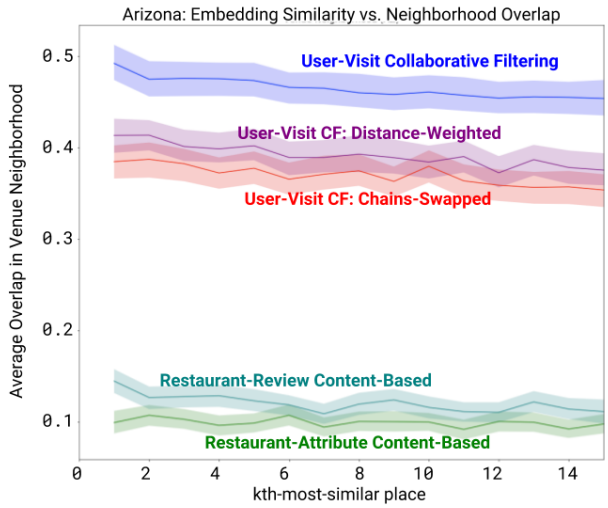


Figure 8.1. Locality Bias for embeddings from Phoenix, AZ, USA metropolitan area.

Locality Bias for embeddings from Phoenix, AZ, USA metropolitan area. Content-based embeddings (Restaurant-Review and Restaurant-Attribute) demonstrate very little correlation between embedding similarity and a restaurant's neighborhood. User-Visit Collaborative-Filtering embeddings demonstrate more positive correlations, though chain-swapping and distance-weighting decreases this locality bias.

Figure 8.2 provides additional evidence of locality bias in user-visit CF embeddings. Within the user-visit CF embeddings, the average cosine similarity between a given location of a chain restaurant and that chain's other locations, which one might expect are nearly identical, is significantly less than the average cosine similarity between a given location of a chain restaurant and a random restaurant from the same neighborhood. This indicates that the user-visit CF embeddings encode location more strongly than content-related concepts such as menu items or price.

8.4.3. RQ2: Effect of User and Place Bridges

In RQ1, we determined that user-visit CF embeddings encode strong locality bias, so we now turn towards understanding how we might overcome that locality bias. We examined two

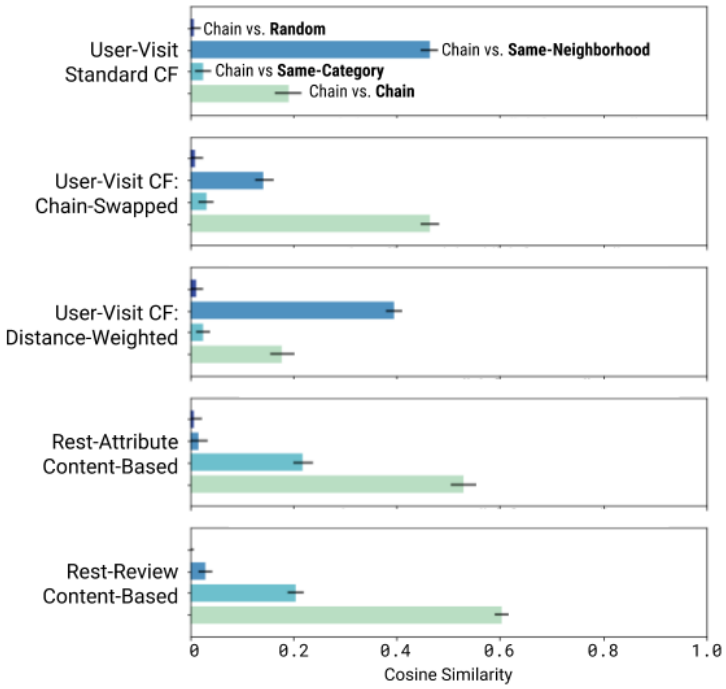


Figure 8.2. Chain-based comparison of locality bias in embeddings.
For each embedding, comparison of the average cosine similarity between a location of a given chain restaurant and other locations of that chain, restaurants in the same neighborhood, and restaurants in the same category. Higher cosine similarity with other restaurants in the neighborhood as compared to other locations of the chain or restaurants in the same category indicates greater locality bias, at the expense of identifying "semantically-similar" restaurants.

potential mechanisms for reducing the locality bias within the embeddings while (hopefully) retaining the valuable signal about people's tastes for restaurants. The first mechanism was through user bridges, or people who visit restaurants across multiple areas. We amplify this signal by recording a visit in the co-occurrence matrix from which the embeddings are learned not as a one for all visits, but as the distance between that restaurant and the user's "home" location or previous visit, whichever distance is shorter. The second mechanism that we tested was through place bridges, or restaurants with multiple locations in a city. For a user's visit to a given chain location, we randomly choose one of the chain's locations to

record in the co-occurrence matrix, reducing the degree to which a user's visits tend to be focused in a physically-constrained region.

We find encouraging results that we can decrease the locality bias of the embeddings. We see in Figures 8.1 and 8.2 that the user bridges and place bridges both significantly decrease the locality bias encoded in the embeddings. The Spearman correlation between neighborhood overlap and embedding similarity reflects this decrease in locality bias, dropping from ρ=0.268 with the user-visit CF embeddings to ρ=0.245 for the chain-swapped embeddings and ρ=0.235 for the distance-weighted embeddings.

8.4.4. RQ3: Bias, Accuracy, and P-fairness

In RQ3, we start by asking whether our interventions to reduce locality bias also reduced the accuracy of place recommendation. We examine two evaluations of the place recommendation model's performance: the traditional approach, which preserves location by evaluating a model within the context of an entire region, and the approach proposed in this work, which controls for location by restricting the model evaluation to a much smaller region containing just 15 restaurants. We posit that performing well on the location-controlled evaluation is the better measure of the efficacy of a place recommendation model. If a model has better performance on the location-preserved evaluation, then that indicates that it has encoded locality bias that does not actually help to distinguish which restaurant best aligns with a user's tastes in an area.

Figure 8.3 shows the recommendation accuracy per each evaluation approach for each of the embedding methods in Phoenix, AZ. We immediately notice the poor performance of the Restaurant-Attribute embeddings, which is not surprising given the simplicity of their

underlying data. We ignore them for the rest of the analysis. In the location-preserving evaluation, we see the standard user-visit CF embeddings perform best, followed by the locality-bias-reduced embeddings and content-based embeddings. In fact, the relative differences in performance match remarkably well with the degree to which each model encodes location (Figure 8.1). These results mirror many studies that find that content-based embeddings are less effective than collaborative-filtering embeddings. In the location-controlled evaluation, however, the other embeddings all perform equally well. This includes the content-based Restaurant-Review embeddings, which is a surprising result. This indicates that the locality bias encoded by the standard user-visit CF embeddings is largely noise in the sense that it is not predictive of what restaurant a user will visit when location is already fixed.

We now explore whether these interventions, by reducing locality bias, led to a more equitable distribution of recommendations across the region—i.e. greater provider fairness. We look at the entropy of recommendations that a model makes sat a given rank across all of the users. We measure entropy in terms of the individual restaurants as well as the neighborhoods where the restaurants are located. High entropy values would indicate that a model recommends many different restaurants across many neighborhoods when its recommendations are viewed in aggregate.

Table 8.4 shows the entropy of recommendations at different ranks. Higher entropy values indicate a more uniform distribution of neighborhoods (first three columns of results) and individual restaurants (second set of columns) in the traditional place recommendation rankings. We see that chain-swapping, despite reducing locality bias, actually leads to a less uniform distribution of restaurants and neighborhoods in its predictions. Distance-weighting leads to a higher entropy (more equitable distribution) in terms of neighborhoods, but lower entropy for individual restaurants. This suggests that while it leads to recommendations

Table 8.4. Provider fairness: entropy of place recommendations.

This table shows the entropy of recommendations at a given rank across all the test users. The entropy is calculated both by neighborhood—i.e. are restaurants from many different neighborhoods recommended or just a few—and restaurant—i.e. are many different restaurants recommended or just a few. Higher entropy values indicate greater distribution and therefore greater provider fairness.

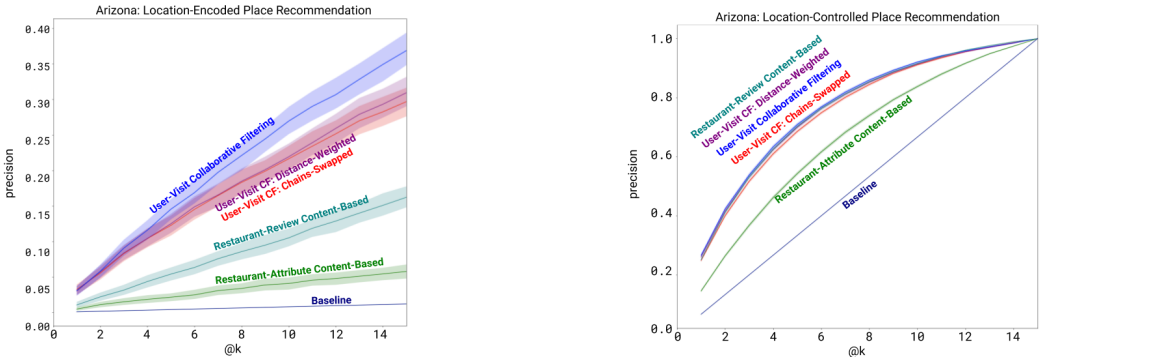
entropy@rank	Neighborhood			All Restaurants		
	Simple	Chain-Swapped	Distance-Weighted	Simple	Chain-Swapped	Distance-Weighted
entropy@1:	3.269	3.007	3.288	5.349	5.139	4.783
entropy@2:	3.369	3.128	3.572	6.082	5.821	5.908
entropy@3:	3.436	3.221	3.604	6.489	6.232	6.276
entropy@4:	3.479	3.28	3.682	6.768	6.505	6.535
entropy@5:	3.503	3.313	3.688	6.97	6.737	6.744
entropy@6:	3.512	3.324	3.667	7.164	6.916	6.894
entropy@7:	3.522	3.34	3.672	7.356	7.095	7.058
entropy@8:	3.533	3.359	3.678	7.506	7.243	7.182
entropy@9:	3.55	3.374	3.652	7.648	7.381	7.303
entropy@10:	3.555	3.387	3.683	7.756	7.499	7.408
entropy@11:	3.558	3.418	3.7	7.866	7.617	7.537
entropy@12:	3.568	3.437	3.695	7.953	7.725	7.631
entropy@13:	3.585	3.43	3.704	8.054	7.803	7.708
entropy@14:	3.583	3.47	3.709	8.124	7.899	7.787
entropy@15:	3.582	3.465	3.696	8.182	7.955	7.841

across a wider range of neighborhoods, it consistently elevates specific restaurants in these neighborhoods. Though these trends hold across the first fifteen ranks, the entropy at the first positions is of utmost importance given that users have a strong bias towards clicking on the top-ranked results [**154**]. These results indicate that reducing locality bias within the embeddings and a given user's rankings does not immediately lead to a greater spatial distribution of recommendations and improved provider-fairness.

8.5. Discussion

8.5.1. Balancing Locality Bias

While it is not clear what magnitude of locality bias is ideal in representations of users or POIs, there are good reasons to not directly encode the locality bias exhibited in past mobility behavior. Additional research should further explore this balance. In this research, we have demonstrated instances in which it may be beneficial to reduce locality bias. The strong



(a) Location-Preserving Evaluation (b) Location-Controlling Evaluation

Figure 8.3. Place recommendation precision for each embedding in Phoenix, AZ. The location-preserving task evaluates a place recommendation model on its ability to rank all restaurants in a large region while the location-controlling task focuses on a small neighborhood. For both settings, the task is to choose which of the restaurants is the held-out visit for a user. Baseline accuracy, the straight diagonal line, is the random chance of ranking the held-out restaurant in the top k places. 99% confidence intervals are shown, but can be very tight.

degree to which user-visit CF encodes a POI's location hindered useful comparisons across large distances and we found that reducing locality bias did not decrease recommendation accuracy. There are also strong arguments that people's spatial patterns are not purely choice, but, at least in the United States, have been driven by governmental policies as well (e.g., [**263, 2**]). Allowing for greater flexibility in place recommendation algorithms, such that do not purely reinforce dominant past mobility behavior, is then an important direction of study. This has been explored in related domains, such as Hirnschall et al. [**145**], who examined how to incentivize Airbnb users to stay at homes with fewer reviews, and Ensign et al. [**80**], who demonstrated that feedback loops in predictive policing due to historical patterns in crime data could direct police resources disproportionately towards certain areas, missing the true geographic spread of crime that occurs in a city.

This work also uncovers a pitfall of bias-reduction techniques. By focusing on reducing locality bias within the embeddings and a user's recommendations, the model did not produce

more equitable recommendations across all users. In fact, the top-ranked restaurants after bias reduction were more concentrated in fewer neighborhoods. This suggests that new techniques may be needed to account for provider-fairness directly.

8.6. Future Work and Limitations

This work provides an initial exploration into how to measure bias within geographic embeddings and reduce locality bias within user-visit collaborative-filtering embeddings. Our dataset was specific to US cities, so expanding to other countries and datasets (e.g., composed of trace data such as phone GPS or click-streams from web queries) would provide insight into how this challenge of locality bias surfaces in other sources of geographic data. Additionally, given the ubiquitousness of map and place recommendation applications (e.g., Yelp), these datasets are likely affected by past implementations of place recommendation algorithms. User experiments would be necessary to truly understand the quality of recommendations. We tested a few simple approaches to learn embeddings, but other approaches, such as graph-based techniques, might introduce different biases.

Further study should expand on the mixed results from debiasing in this work. Better results might be achieved through adapting more advanced techniques for debiasing [**29**] or adversarial objectives [**248**] to removing locality bias in place recommendation embeddings. Additionally, techniques that promote diversity in rankings (e.g., [**231**]) should be explored to determine whether they might improve provider-fairness.

8.7. Conclusion

We explore locality bias in user-visit collaborative filtering, a common approach to personalized place recommendation models. We find that user-visit collaborative filtering approaches strongly encode physical location when compared to content-based approaches. While this locality bias derives from the nature of human mobility data and is often a core aspect of place recommendation, it is also not clearly a desirable feature as it encodes existing patterns and reduces the effectiveness of comparing users and restaurants separated by larger distances. We introduce two techniques for controlling this locality bias: user and place bridges. Within a city, amplifying users who bridge neighborhoods (i.e. users who further from home) and places that bridge neighborhoods (i.e. chains) reduces this locality bias while maintaining recommendation accuracy. We find, however, that these locality-bias-reduction techniques do not lead to greater provider-fairness. That is, the distribution of restaurants and neighborhoods in the top-most recommendations across all users actually is less equitable after applying these corrections.

CHAPTER 9

Geographic Embeddings

In this chapter, we adapt techniques from natural language processing to support geographic representation learning and explore how we might evaluate the bias encoded in the resulting embeddings.[1]

9.1. Introduction

There is an acknowledged disconnect between, on one hand, the rich literature in the field of human geography on how people and communities vary in their characteristics across space, and, on the other hand, how we often operationalize geography in spatial computer modeling and prediction [**8, 220, 281, 110**]. Representing places (e.g. a city) is an important consideration because the characteristics of a place are critical context in understanding and predicting how individuals will behave—e.g., how people vote [**100, 159, 287**], where they might visit [**22, 233, 331**], adoption of ride-sharing services [**47, 259, 301**].

Approaches to statistical modeling of spatial processes have largely represented geography at its simplest—i.e. as space with distance determining the similarity of two places—and ignored the richness of details that determine how related one region is to another—i.e. place. However, when attempts are made to extend representations of geography beyond distance (e.g. by incorporating in population demographic variables), the models often are complex, constrained to a scale and region that has the requisite data, and therefore do not generalize outside the specific bounds of the study. Determining the most effective variables

[1]The work presented in this chapter builds on research in collaboration with Brent Hecht and Shilad Sen.

and scale at which to study a phenomenon (e.g. via grid search) is often intractable due to the complexity of point-in-polygon or nearest-neighbor operations.

Recently, there has been an exploration of methods that go beyond distance and population demographics and seek to take advantage of the large volumes of (open) geographic data to help to characterize the complex ties between different places—e.g., social media such as photos [**251, 270**], tags [**22, 162, 252**], or location check-ins [**55, 59, 185, 309, 349**], peer-produced content such as OpenStreetMap [**59, 212, 309**], Yelp [**8**], and Wikipedia [**272**], sensor data such as mobile phone metadata [**28, 59**], satellite imagery [**152, 297**], or Google Street View imagery [**69, 100, 226**]. This body of research has demonstrated the promise of large, noisy geographic datasets to provide useful signals in specific geographic problems such as POI recommendation, predicting demographic attributes, learning flexible regions, or determining the similarity of places.

Furthermore, by departing from sources such as census statistics that may be infrequent or highly aggregated to trace data that is more real-time and fine-grained, these approaches have shown promise for bringing the benefits of research and internet technologies to these regions that traditionally have lacked data—e.g., many countries in Africa [**28, 69, 152**]. Assuming that these models are carefully developed, this is a valuable opportunity to help reduce the burden of contribution in these regions as opposed to merely reflect existing structural inequalities.

There is tension, though, between what patterns a designer intends a model to learn, and the patterns that the model actually learns. That is, are these new sources of data truly replacements for more structured and direct measurements like censuses or rather fragile or problematic proxies for the patterns that we actually seek to predict? An illustrative example is Penny, an AI system trained to predict a neighborhood's median income from

satellite imagery [**26**]. Penny was designed to be interactive and encourage users to explore their and the models' biases (the online platform allows users to add elements to a region and see how the predicted income changes) [**26**]. News coverage of the platform discusses limitations such as correlation does not imply causation and the the lack of explanations for the predictions, but a glance at the headlines (e.g., "An AI That Predicts a Neighborhoods Wealth From Space" [**108**], "A Mapping Machine Identifies Wealth From Space" [**26**]) reveals that it is accepted that the model has learned to identify wealth from satellite imagery. While predicting wealth is the goal, the potential biases or other patterns that the model might have learned and end up reinforcing if this type of model was actually used to track wealth (e.g., [**152**]) would appear manifold. For example, there are very strong correlations between wealth and race in the US [**174**] and a long history of residential segregation (see § 2.2), separate models for Penny had to be trained for each city [**108**] and related models have found that many of these signals are only robust predictors in their local context [**320**]. If Penny actually encoded race as opposed to income, this could result in a greater burden being placed on minority communities to prove their wealth, which would be problematic for contexts like predicting the value of homes[2] or economic opportunity [**149**]. It is not clear, however, how one might evaluate to what degree Penny, and related models, may be encoding these biases.

The goal of the research contained within this chapter is to help balance 1) the general promise of using geographic trace data to build useful predictive models, with, 2) the challenge of understanding what additional patterns are encoded in these models. To satisfy these two goals, we develop general methods for learning from geographic trace data that also afford direct evaluation of their biases. Specifically, we build on past work that uses

[2]https://www.kaggle.com/c/zillow-prize-1

open geographic data, but focus on the challenge of learning *general* representations—i.e. embeddings—of places from this data. While incorporating open geographic data directly to solve specific tasks (as outlined above) will generally lead to higher performance than building general embeddings as an intermediate step, general embeddings have the following important properties that support both of our goals:

- General embeddings provide a unique tool for exploring the nature of large datasets (e.g. biases) and developing methods for removing some of these negative aspects. That is, they are highly extensible.
- Openly-available general embeddings can greatly reduce obstacles to incorporating geographic context into spatial models and allow fine-tuning from smaller datasets without overfitting. Many machine learning frameworks are developed to incorporate embeddings as a standard input.
- With careful choices regarding a minimum level of aggregation, general embeddings learned from sensitive data (e.g. mobile phone location traces, which might be the only or most representative data available for a region [**28**] can still be distributed openly.

These benefits have been validated in the natural-language and image-processing communities. Openly-available, pre-trained general representations of words (e.g. Google News embeddings [**219**] have provided an incredibly valuable tool for tackling complex challenges such as image captioning (e.g. [**165**]), making natural language processing tutorials far more accessible (e.g. [**46**]), and understanding the nature of different text corpora (e.g. [**16, 29, 63**]). Likewise, openly-available, pre-trained image processing models (e.g. VGG16, which was trained on ImageNet [**288**]) have allowed researchers to focus on specific image processing

tasks where there is much less supervised data available (e.g., economic prediction from satellite imagery in African countries [**152**]). These benefits have not been without drawbacks though: word embeddings have been widely shown to encode gender biases [**29, 38**], which then are incorporated into downstream technologies like machine translation technologies until platforms acknowledge the issue and develop solutions that account for these biases—e.g., Google Translate [**158**].

Specifically, in this chapter, we adapt natural language processing methods to learn representations of place (geographic embeddings) from two different datasets of geographic traces: 1) OpenStreetMap tags and 2) Flickr photos. user tags from Flickr photos. We explore how findings from prior chapters can guide choices of *geographic hyperparameters* (see §2.5) in the preprocessing and learning algorithms.

We evaluate what is encoded within these geographic embeddings by evaluating their effectiveness at predicting various demographic attributes. We find that:

- Embeddings learn sufficient information: all are better than random but generally fall short of the amount of content encoded by three census attributes.
- Race is an outlier: the embeddings encode more information about race than auxiliary census attributes.
- Generalization: While OSM tags perform best under standard testing where census tracts are randomly held-out, there is some evidence that Flickr tags may best generalize when entire regions are held-out.

Finally, given that these geographic embeddings represent a novel way of encoding and understanding geography, we discuss their limitations as well as how concepts from the

geography and machine learning communities might further extend their development and potential uses.

9.2. Related Work

On top of the extensive background contained in Section 2.5, we provide more context about representation learning below.

9.2.1. Representation Learning

Over the last several years, large gains have been made in natural language modeling in part due to improvements in the ability to learn general, unsupervised representations of complex entities such as words—e.g. skip-gram modeling [**219**]) as vectors that can be efficiently fed into neural networks [**23**]. That any individual who is building language models can start with a pre-trained set of representations (e.g., word embeddings) greatly reduces the volume of data (and therefore training costs) necessary to achieve high accuracies in other language tasks such as word sense disambiguation [**148, 165**].

Within geography, there is no common set of place representations or method of learning embeddings that are available to researchers or practitioners. Researchers have focused on specific tasks such as applying clustering techniques to learn unsupervised regions (e.g. [**55, 340, 162**]), sequence modeling techniques such as word2vec to place visits to support place recommendation tasks [**83**], and image-processing techniques to satellite imagery to predict human attributes (e.g. [**152**]). We seek to extend representation learning to geography such that the resultant embeddings can be readily incorporated into a wide variety of geographic prediction tasks.

Table 9.1. Representation learning datasets.

Descriptive statistics of each dataset from which geographic embeddings were learned.

Dataset	Filtering	# Unique Tags	# Total Tags
OpenStreetMap	Default (all tags)	7,034,615	40,982,241
OpenStreetMap	Filtered	119,207	15,384,843
Flickr YFCC100M	User tags	461,189	23,828,323
Flickr YFCC100M	Machine-learned tags	1,570	21,457,049

9.3. Data and Methods

Below we motivate and describe the choices that we made in constructing the pipeline through which we learned and evaluated our geographic embeddings. Just as machine learning algorithms often have important hyperparameters that determine their efficacy (e.g., learning rates, regularization), we have *geographic* hyperparameters, which are design decisions that directly relate to the how a place is represented in the final embeddings. Geographic hyperparameters present a unique challenge in adapting methods developed for language corpora to a geographic context. We describe these choices as well as the representation learning algorithm that we adapted for use and how we evaluate the impact of these choices through a demographic prediction task.

9.3.1. Data Sources

Perhaps the most important geographic hyperparameter when building embeddings is the choice of dataset. The resulting embeddings, as a dimensionality reduction, will encode the biases contained within the dataset. For an overview of potential choices, see Table 2.1. In this work, we build embeddings from two very different sources of geographic data: OpenStreetMap and Flickr. Details of each dataset are provided in Table 9.1.

9.3.1.1. OpenStreetMap. OpenStreetMap (OSM) is an open, crowd-sourced map of the world that contains information about the physical world (e.g. locations of administrative boundaries, buildings, roads, trees) as well as structured tags that describe these entities (e.g. addresses, business types, road speeds, tree species) in the form of key:value pairs like "amenity:church". Because OSM is an openly-editable platform, it accepts all types of data contributions with minimal guidelines as to what information is expected. OpenStreetMap tags should provide a rich source of data for learning representations of places based on what exists there and the attributes of these objects. They also present many challenges though. For instance, OSM has accepted bulk uploads of governmental datasets throughout its history and thus much of the tag data pertains to metadata that has a more tenuous connection to geography (e.g. upload dates, information source). Additional challenges include variable coverage [**275**], propensity towards bot-generated content in rural areas (§4), varying conventions about how to select tags [**123**], and other biases that commonly appear in online systems [**293**].

Given these data quality challenges, we explore two sets of tags from OpenStreetMap: 1) a *default* approach where no tags are filtered out with the intuition that the default parameters for the representation learning algorithm (described below and uses negative sampling, which downplays the importance of high-frequency tokens) would be sufficient to filter out much of the noise, and, 2) a *theory-motivated* approach where one author went through the 1000 most common key values in OpenStreetMap tags for North America (representing 73% of total tags) and removed any tags that did not refer to general characteristics of places (i.e. removed tags such as place names, administrative codes, miscellaneous identifiers that would only have a coincidental tie to geography such as upload dates or data sources). The remaining tags encode both aspects of physical geography (e.g. information about land

cover, bodies of water, trees) and human geography (e.g. types of roads, businesses, zoning), which are important components of place [**25**]. While there is a degree of arbitrariness in this procedure, it reflects a first-principles and best-effort approach to remove noise and only include information that should reasonably inform the paragraph-vector model about the nature of a given place.

9.3.1.2. Flickr. In order to better understand the effects of selecting OSM tags, a peer-produced dataset, we also examine Flickr tags. Flickr is a social media photos platform. We use a public dataset of Flickr photos (YFCC100M [**302**]), selecting just the photos that are geolocated to the regions for which we build embeddings. Each photo is labeled with two types of tags that we consider: user-provided tags, which are unstructured but potentially provide much more insight into how people view a place, and automatically-generated tags (and probabilities) from a deep-learning AlexNet model fine-tuned to classify each photo based on 1570 different concepts (e.g., architecture, lake, quirky). Though both tag datasets reflect the same underlying photographs, they represent two very different means of representing them. As with OpenStreetMap, we build separate embeddings for each type of tag.

9.3.2. Unit of Aggregation

A major decision made in spatial modeling is the scale at which the data will be represented. As discussed in Section 2.5.3, different units encode different burdens of producing content and map better or worse to the spatial process being studied. We can choose between administrative units, which generally better reflect human geography, and grid systems, which are much more computationally efficient but ignore political and social boundaries.

We can also choose the scale at which we aggregate the data. In this work, we implement and compare both options: a grid-based system and administrative units.

For the grid-based system, we use OpenLocationCode (OLC),[3] a hierarchical grid system with constant-time operations for determining which grid cell contains a given latitude-longitude coordinate. The first five levels of the original OLC hierarchy have grid cells with lengths of approximately 2200km, 110km, 5.5km, 275m, and 14m. The OLC scales were designed to simplify address geocoding though, not to optimally model human geography.[4]

For the administrative units, we select census tracts. Census tracts were chosen for several reasons: 1) they are the smallest-level units for which there is comprehensive census data available (necessary for the task described below), 2) they are relatively consistent in population, which ensures a more even amount of content for each census tract, and, 3) they are of a size that tends to capture residential segregation dynamics per [**286**], which is a strong motivation for understanding how geographic trace data captures potentially undesirable patterns.

9.3.3. Representation Learning Algorithm

We adapted methods from the natural language understanding community to learn our geographic embeddings. Given that all of our datasets are in the form of tags that are associated with a given spatial unit, we can apply a consistent representation learning algorithm to each dataset. Specifically, we use Paragraph-Vector (or doc2vec) [**184**], with adaptations as described in this section to make it more suitable to the geographic realm.[5]

[3]http://openlocationcode.com/

[4]To capture a wider range of scales, we also redesigned and tested an OLC hierarchy that only scaled by a factor of 5 at each step (as opposed to 20). Specifically, the grid cells in our version lengths of 625km, 125km, 25km, 5km, 1km, and 200m. We found minimal difference.

[5]We also tested LDA models but found that they performed less well.

Le et al. [**184**] introduce a technique to learn vector representations for variable-length documents—i.e. not just a single word as in word2vec [**219**] but a collection of words that together define a document such as a movie review. Paragraph-vector is an unsupervised technique where the model learns a fixed-length vector representation of a given "document" iteratively through backpropagation by attempting to predict words contained within that document using that document's representation and other words in the document as context.

We adapt this technique to each of our four datasets of tags by collecting all of the tags from that dataset that are contained within or intersect a given geographic boundary. For OpenStreetMap, tags can be applied to nodes (points), ways (lines), and relations, and we include tags from all of these geometries. For Flickr, tags are associated with a photo with specific coordinates, making the calculation of intersection trivial. These tags are the equivalent of the words in paragraph-vector and the geographic boundary of the place is equivalent to the document. A vector embedding is then learned (we used the implementation provided by genism [**357**] and set the embedding length to 64) for that area.

In our implementation, we largely depended on the default hyperparameter values from Le et al. [**184**]. Specifically, we use the distributed bag-of-words (DBOW) implementation (word-order does not clearly apply to a collection of tags) with the following hyperparameters: negative sampling (5 samples drawn), down-sampling of words with a frequency greater than 0.001, 4 iterations, a minimum count of 5 for a given tag key or value to be included, initial learning rate of 0.1 dropping to a minimum of 0.0001.

A useful side-effect of paragraph-vector is that, during training, it also learns embeddings for the words that define the documents (i.e. OpenStreetMap tags). These tag embeddings provide insight into the effectiveness of this technique for learning a notion of similarity

between geographic entities. For example, the most similar tag to "Burger King" (a fast-food burger chain) is "McDonald's" (a very similar fast-food burger chain) as one might expect. However, because this notion of similarity is a largely function of geographic co-occurrence (appearing in the same spatial unit), Burger King also has a higher cosine-similarity to "CVS" (a pharmacy) than "In-N-Out" (a fast-food burger chain, but one that is trendier). In this case, CVS is learned to be more similar to Burger King because it is common for strip malls or other commercial districts with Burger King restaurants. We further quantitatively test the quality of the OpenStreetMap tag embeddings by comparing the cosine similarities between them with a crowd-sourced ground-truth dataset of semantic similarity between OpenStreetMap concepts [**15**] (no equivalent task exists for Flickr tags). We achieve a Pearson correlation coefficient of 0.426 with the unfiltered OSM census-tract embeddings, which is lower than the performance achieved by their ensemble of natural-language models ($r = 0.737$) [**15**], and the performance of a domain-specific geographic semantic relatedness algorithm trained on a combination of natural language and geographic signals ($r = 0.813$) [**272**]. Given that it is an unsupervised method that is based solely on spatial co-occurrence, it is not a poor result.

9.3.4. Tasks

We selected the task of census prediction for evaluating the resulting geographic embeddings. In this task, a model is built to predict the demographic data for a given census unit. For example, embeddings can be learned for census tracts and then an appropriate model can be trained that predicts the census demographic for the census tracts from their embeddings. This type of task has been undertaken in the past through a variety of methods such as via satellite imagery [**152**], mobile phone data [**28**], and Google Street View Imagery [**100, 106**].

Table 9.2. Demographic Variables

Categories of each ACS 2016 5-Year variables used in census prediction task. The categories for each demographic are represented by proportions of the population and sum to one in all cases, so the models used in predicting them all used a softmax activation function as the final layer.

Demographic	ACS Categories
Race	White; Black or African American; Asian; Some other race; Two or more races
Income	$1 - $9,999 or loss; 10,000 to $14,999; 15,000 to $24,999; 25,000 to $34,999; 35,000 to $49,999; 50,000 to $64,999; 65,000 to $74,999; 75,000 or more
Education	Less than high school graduate; High school graduate; Some college; Bachelor's degree; Graduate or professional degree
Age (years)	Under 5; 5 to 17; 18 to 24; 25 to 44; 45 to 54; 55 to 64; 65 to 74; 75 and over

Specifically, we learn embeddings for census tracts across 14 different counties comprising 8 cities: Boston, MA, New York City, NY (5 counties), Chicago, IL, Twin Cities, MN (2 counties), San Francisco, CA, Los Angeles, CA, New Orleans, LA, Houston, TX. For each census tract, we download the 2016 5-year American Community Survey (ACS) data from Table S0601 and extract variables related to race, educational attainment, income, and age (see Table 9.2 for specifics).

For each demographic category (race, education, income, age), we train a neural network to predict the attribute and evaluate their performance via 10-fold cross-validation until the accuracy on a held-out validation set did not increase for three epochs. For the census tracts, there is a direct mapping between the embedding and demographic information. For the grid cells, we identify the census tract that contains the center of the grid cell and use the demographics from that census tract. KullbackLeibler divergence is used for the loss metric as each demographic category is represented by a set of buckets and corresponding proportions of the population for that census tract that sum to one. The models are evaluated through computing the Spearman correlation between the actual proportions and predicted proportions for each demographic bucket.[6]

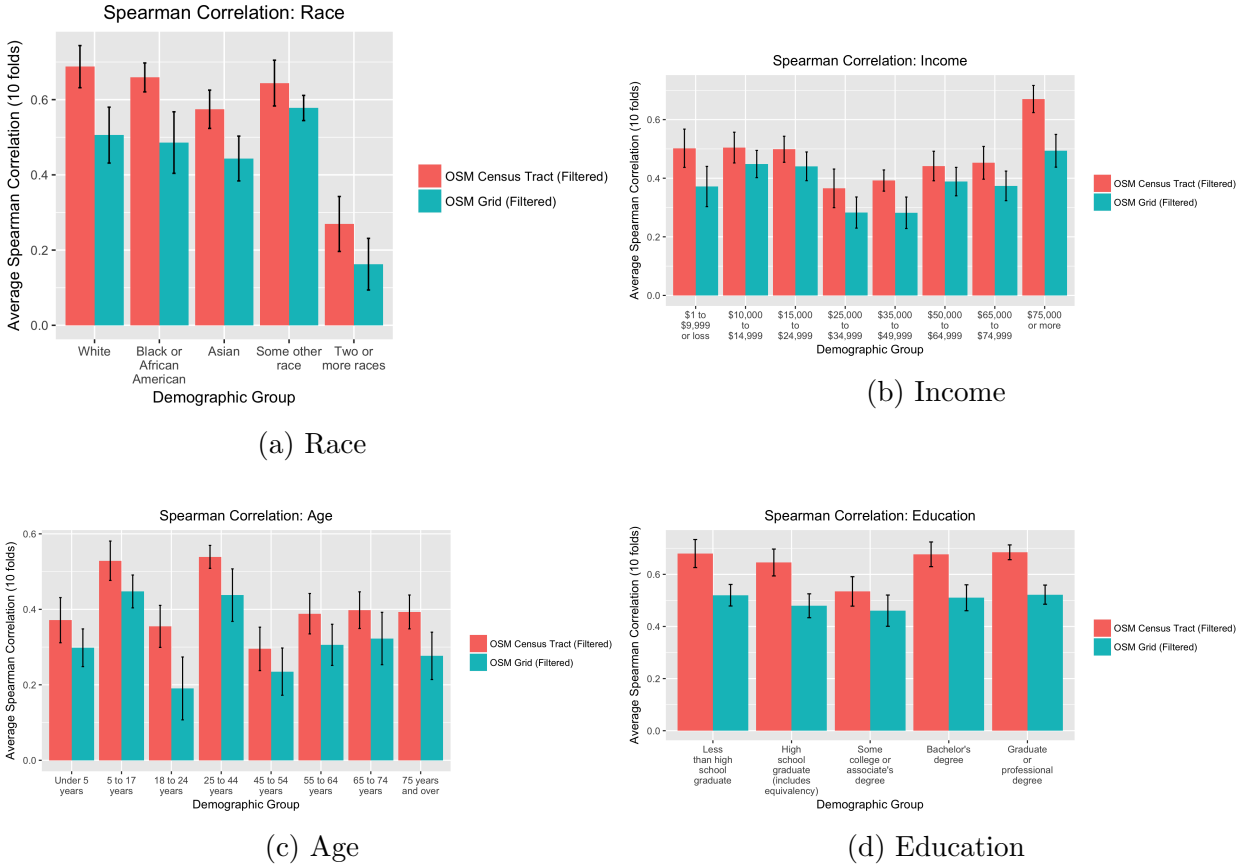
[6]Absolute error was also examined and produced the same results. Spearman correlation is reported as it captures how well the model can rank the held-out census tracts by a given demographic—e.g., highest proportion Asian to lowest proportion Asian.

We explore two methods of splitting training/validation/test sets. The first is a random split 80:10:10 across all spatial units. This results in each city having proportionate numbers of spatial units in the training, validation, and test sets. Given that geographic prediction models often struggle to generalize to new areas (e.g., [**66, 226**]), we also experiment with holding out a state at a time—e.g., train/validate on Boston, New York City, Houston, Twin Cities, New Orleans, Houston; test on San Francisco and Los Angeles. Performance on this evaluation should give a better indication of how well a given embedding generalizes.

9.3.5. Baselines

In order to evaluate how strongly the embeddings were encoding various demographics, we need interpretable baselines against which to compare. Specifically, we compare the results from the embeddings models (average of the five folds) against the following baselines:

- Random (lower bound): predictions made with an analogous model that takes as input embeddings that are just random noise, which should approximate the intercept or the average of the training data for that demographic. This would indicate that the embeddings hold no information about that demographic.
- Census (middle bound): predictions made with an analogous model that took as input the values for the three other ACS demographics—-e.g., for predicting race, a model was built that took as input that census tract's age, income, and education values.
- Neighbors (upper bound): the median of the values from that census tract's neighbors. Due to segregation and spatial homophily, this heuristic is actually quite effective. Notably, this baseline is not possible for the held-out state evaluation because each unit's neighbors are also held-out.



(a) Race

(b) Income

(c) Age

(d) Education

Figure 9.1. Each figure contains the average Spearman correlation for a given model for that demographic variable. 95% confidence intervals given as calculated by mean and standard deviation across 10 folds.

9.4. Results

We walk through the comparisons for each geographic hyperparameter below. We begin with aggregation unit as it allows us to narrow down the rest of the comparisons. Higher Spearman correlations indicate that a given embedding encodes more information about the area—i.e. will likely achieve greater performance in a predictive model—but also that the embedding is encoding more information about the demographics of the individuals who live in that region.

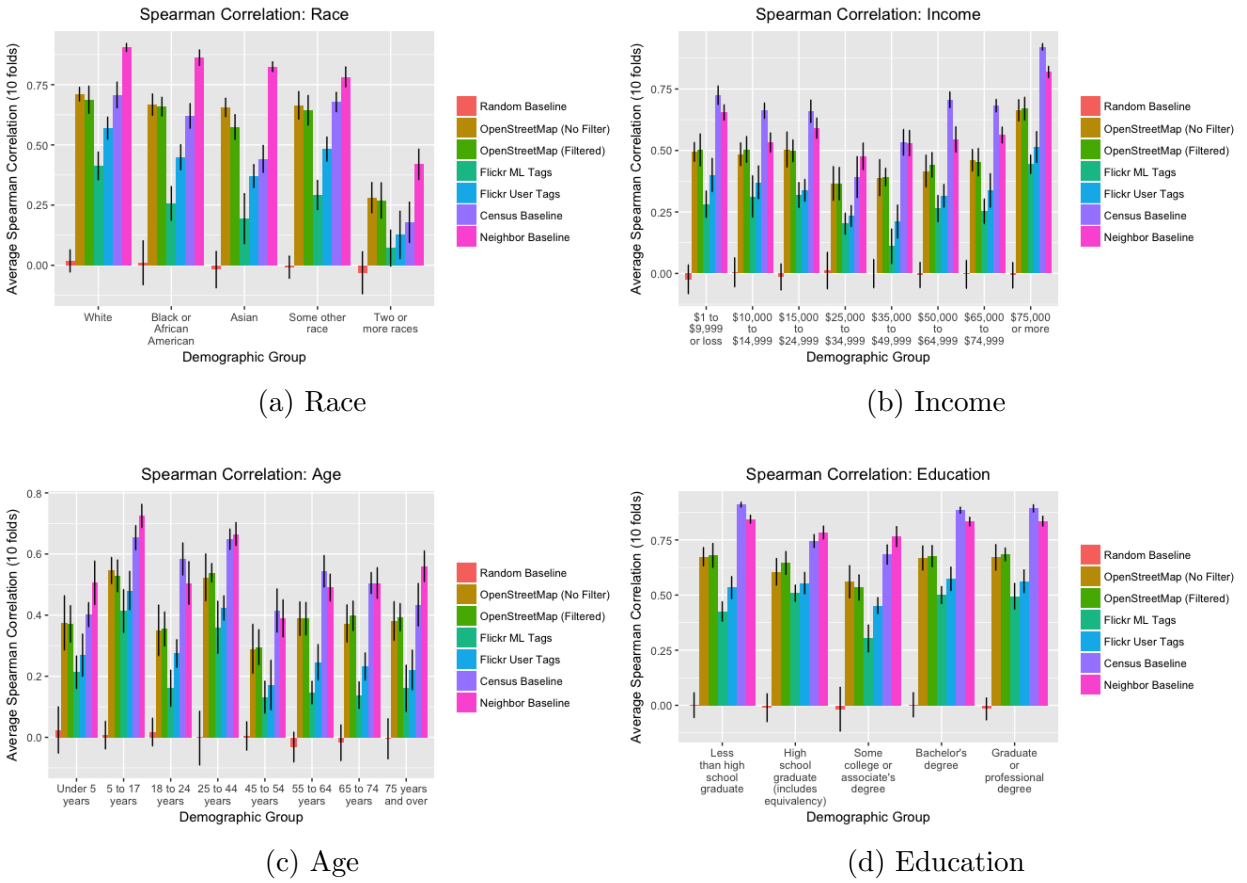
9.4.1. Aggregation Unit

Figure 9.1 shows the results for the comparison for embeddings aggregated to grid cells (OpenLocationCode) and administrative units (census tracts). It focuses on the results for the OpenStreetMap filtered tags and all other hyperparameters are kept constant between the models. The census tracts consistently have higher Spearman correlations, indicating that the census tract boundaries are more effectively capturing aspects of place than the grid cells used here. We will focus on the census tract results for the rest of this section.

9.4.2. Data Source

Figure 9.2 shows the results for each of the demographic categories using the embeddings aggregated to census tracts and based on random training/validation/test sets. Each of the three baselines are included in the figures to provide context for how strongly a given embedding is encoding a given demographic variable.

A few general trends emerge. The performance of the OSM and Flickr embeddings substantially out-perform the random baseline (equivalent of no information encoded). This indicates that the general embeddings strategy is successfully capturing systematic differences between different census tracts. The performance of the each embedding, however, is generally significantly lower than both the census baseline (information encoded is the equivalent of the other three demographic categories considered) and neighbor baseline (information encoded is the equivalent of averaging a given census tract's neighbors). For race and age, the neighbor baseline encodes the most information. For education and income, the census baseline encodes the greatest information. These results do not show why this



(a) Race (b) Income

(c) Age (d) Education

Figure 9.2. Each figure contains the average Spearman correlation for a given model for that demographic variable. 95% confidence intervals given as calculated by mean and standard deviation across 10 folds.

is true, but it would seem that it is some mixture of residential segregation more strongly encoding race and age while other demographic features not being as highly correlated.

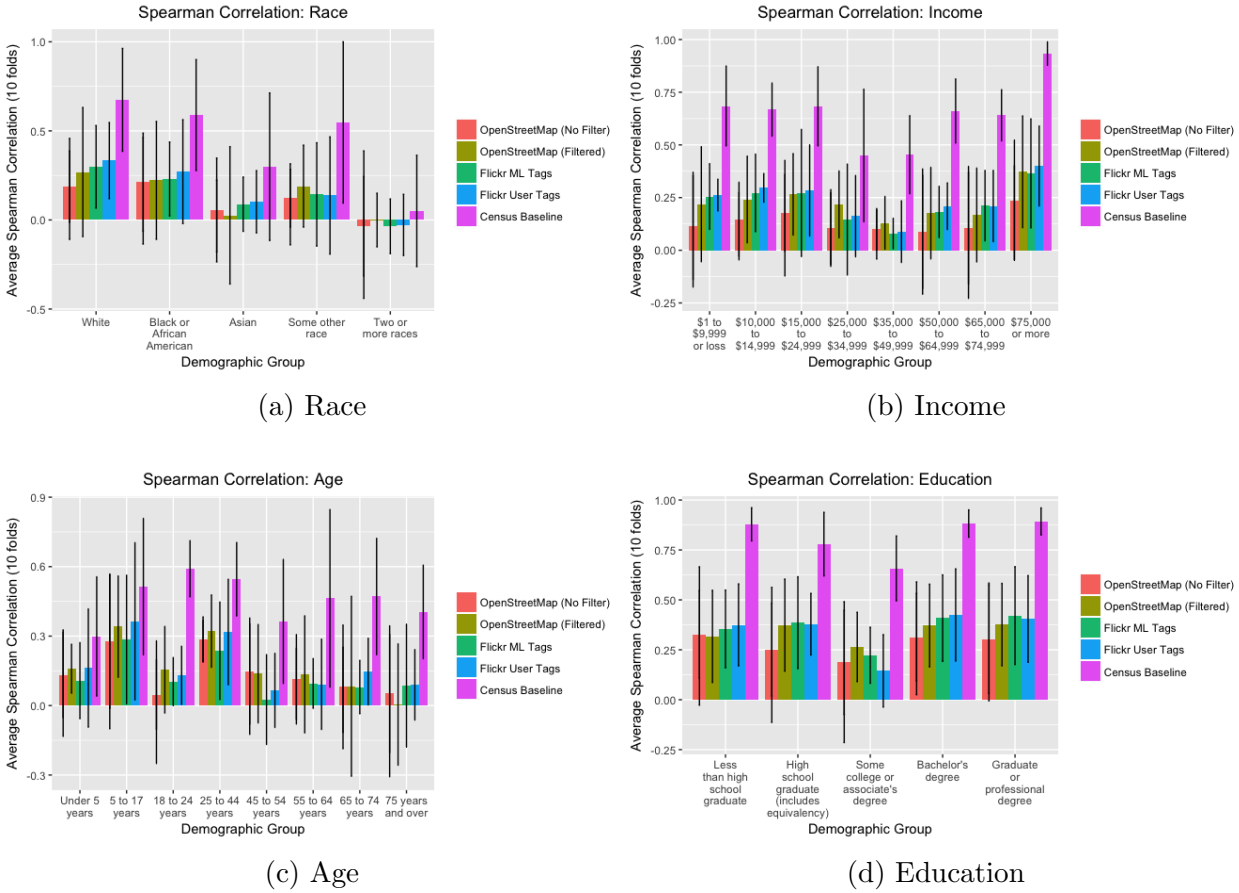
Within the geographic embeddings, we tend to see the highest to lowest performance in this order: OpenStreetMap (No Filter) and OpenStreetMap (Filtered) are generally equivalent followed by the Flickr User Tags and then Flickr Machine-Learned Tags. This is a trend we return to later as it is highly dependent on the evaluation. Specifically, this trend holds when random census tracts are held-out, which means each city has proportionately

the same amount of training data and every census tract in the held-out set likely has a neighboring census tract that was included in training.

There are also a few notable exceptions in the results. For race (Fig. 9.2a), we see that both OpenStreetMap embeddings more strongly encode race than the census baseline. While there is no significant difference at 95% confidence between the three three embeddings (OSM and census) for predicting the proportion of the population that is "White" or "Some Other Race", the OSM embeddings both significantly outperform the census baseline for "Asian" and "Two or more races". For "Black or African-American", there is some indication that the embeddings outperform the census baseline but the difference is not significant. For "Asian" in particular, the OpenStreetMap embeddings substantially out-perform the other embeddings and census baseline.

9.4.3. Type of Evaluation

Figure 9.3 is the equivalent of Figure 9.2 as discussed above but using a held-out state as evaluation. Comparing the two, we can immediately see that the average performance of the embeddings drops substantially and the variance of performance greatly increases. Though there are not significant differences in the different embeddings, we see a reversal of the general trend of OpenStreetMap outperforming Flickr user embeddings in turn outperforming Flick machine-learned embeddings. That is, the embeddings that performed worse in the random held-out evaluation show some indication of performing best when being forced to generalize to new regions.



(a) Race

(b) Income

(c) Age

(d) Education

Figure 9.3. Each figure contains the average Spearman correlation for a given model for that demographic variable. 95% confidence intervals given as calculated by mean and standard deviation across 10 folds. Test sets are held-out states as opposed to random splits.

9.5. Discussion

9.5.1. General Geographic Embeddings

One takeaway from this work is that unsupervised general embeddings of place do successfully encode information that can be harnessed by predictive models. Adapting methods from natural language processing to geography was sufficient for at least one task: predicting the demographics of the population who live in a region. There is certainly still much tuning and

development of methods to achieve comparable utility to word embeddings, and this work provides a groundwork for that continued development. Different forms of geographic trace data will require alternative approaches (see Table 2.1), and hopefully adapting methods from other fields such as image processing will also prove effective.

9.5.2. Evaluating Embeddings

This work explores the tension between viewing effective performance of embeddings in predictive tasks as both a marker of quality and potential liability through encoding potentially spurious and damaging relationships. It sought to bring some transparency to the question of what information is encoded is large-scale geographic trace data. There are some takeaways from the experiments that demonstrate both potential pitfalls and potential approaches that might yield more effective embeddings.

For the random-split evaluation, the OpenStreetMap embeddings actually encoded race more strongly than additional census demographics—i.e. income, age, and education data as per Table 9.2. This is worrisome in that models that replace more transparent information like information about a neighborhood's demographics with geographic trace data may very well end up encoding race more strongly than before.

The experiments also provided indications of how to build more robust embeddings that would hopefully be more transparent as well. While OpenStreetMap, which has significantly more tags and unique tags, outperformed Flickr in the random-split evaluations, this trend reversed in the heldout-state evaluation. This suggests that the OpenStreetMap embeddings learn highly local tags that relate nearby neighborhoods, effectively encoding residential segregation. The Flickr machine-learned tags, which had many fewer unique values but overlapped more across cities, generalized more effectively. This warns against building

geographic prediction models with trace data that has highly local patterns as this will not force the model to learn general patterns. Having fewer unique tags from which the embeddings are learned also makes inspecting and explaining the underlying data much simpler.

9.6. Future Work and Limitations

This project is a start to exploring bias and means of controlling for bias with geographic algorithms (and representation learning more specifically). Some future work and limitations are discussed below.

9.6.1. Towards General Embeddings

This work began the exploration of general geographic embeddings. This approach was undertaken primarily because it affords the evaluation of what is encoded in large-scale geographic data. General embeddings also bring a lot of promise though, especially in supporting the inclusion of data-poor regions in modeling. Expansion should explore additional approaches, more regions, and new data sources. Ideally, future work will also follow several guiding principles that informed the choices in this work:

- Global coverage: avoid datasets that are specific to only a few countries, which would limit the scalability of these methods to areas that already often lack rich geographic data (e.g. census data).
- Open datasets: use publicly-available data to improve the replicability of the work. Eventually when methods are more refined, publishing embeddings based on proprietary data may allow for dissemination of data that otherwise would be too private to share but that contains rich signals regarding place.

- Flexibility: starting from point data as opposed to areal data allows for aggregation to any scale or geographic boundary as appropriate.
- Variety of geographic signals: selecting datasets that provide different perspectives on what comprises place—e.g. both physical and human geography—may lead to more robust embeddings. It is possible that separating these two signals and building separate embeddings would also allow engineers and researchers to evaluate and tune for their relative importance for various prediction tasks.

9.6.2. Reducing Demographic Signals

Looking towards the natural language processing field, an approach that has been tested for removing biases such as those associated with gender from word embeddings is debiasing (see [**18**]). For debiasing, the goal is to learn the desirable features while removing problematic associations encoded within the embeddings. One approach to this is through adversarial learning objectives [**248**], which would allow, for example, one to learn an embedding that encodes features such as income while explicitly not encoding race. Given the systematic nature of residential segregation, this very well may not be possible, but future work should explore this challenge. Even asserting that it is impossible to effectively achieve both goals provided above would be useful for guiding development of geographic predictive models.

9.7. Conclusion

We explore how to build general geographic embeddings from large-scale geographic trace data: OpenStreetMap and Flickr tags. We demonstrate how choices about geographic hyperparameters associated with the representation learning affect the resulting quality of

the embeddings. The quality of embeddings is evaluated in a way that explicitly discusses the tension between high performance and encoding of problematic demographic associations. We show that embeddings that appear to be effective might actually be encoding hyper-local signals that do not generalize well and that strongly encode race compared to other features such as age, education, or income.

CHAPTER 10

Conclusion and Future Work

Across the six studies described within, this thesis presents a number of best practices and lessons learned from studying geographic user-generated content and algorithms. Each individual chapter presents recommendations specific to the domain under study but there are also a number of higher-level contributions that this dissertation makes. We strive to not base any conclusions off of a single data point, city, or platform to ensure the robustness of these contributions. The majority of this work focuses on the United States, as the region in which I am most aware of the history and context, but we also include China (§4) as well as London and Manila (§7). We evaluate a wide variety of platforms across social media—Foursquare (§3), Twitter (§§3,5,7), Flickr (§§3,7,9)—peer production—Wikipedia (§4), OpenStreetMap (§§4,9)—and online services—Google Maps (§7), Mapquest (§7), Yelp (§8). Research that I was involved with and informs this work, but is not explicitly described here, has also studied geographic biases in Airbnb and Couchsurfing [**170**], cross-platform relationships in Google Search, Wikipedia, Reddit, and StackOverflow [**311, 214**], and algorithmic bias in sentiment analysis algorithms [**63**]. Together, these studies consider the questions of inequality (or alternatively, bias or fairness) across many different contexts.

10.1. On Determining the "Representativeness" of UGC

The main contribution from Part I is demonstrating that *equal participation across online communities does not ensure equal benefits from these platforms.* We reach this conclusion by

quantifying online representation in the context of the end consumer of the content. Much of the literature and discussion about the representativeness of online platforms (see §2.3) had been focused on participation rates (i.e. differences in the proportion of demographic groups that use a given online platform) or coverage (i.e. amount of content produced in a given area or about a given topic). As a result, achieving equal representation was often framed as achieving equal participation rates or coverage across all groups. While a random sample would then be representative of the general populace, this proposed end-state ignores other structural biases associated with online platforms (see [**155**]), whose importance can only be understood in the context of how the data is being consumed. Acknowledging the barriers to equal online representation motivates the need for algorithms and evaluations that are more flexible to these realities.

For instance, despite the promise of the World Wide Web as the "great equalizer", we empirically demonstrated that online platforms have largely preserved structural inequalities such as population density in who receives the benefits of this user-generated content. In Section 4, we found that the practice of each town in the United States having a separate Wikipedia article has preserved the low population density of rural areas. This has resulted in a large per-capita burden to generate content in rural areas and generally much lower quality articles. In Section 3, we saw that using counties to aggregate content (as a spatial unit that preserves population density) results in a lower proportion of content being produced by local individuals and often insufficient content to be included in social science research. In Section 5, we saw this same pattern for geolocation inference algorithms. All three of these findings demonstrated that equal participation alone would not lead to equal benefits across urban and rural areas.

We did not explicitly test how to overcome this challenge of population density in these systems, but this work does point towards some potential solutions. In Section 2.5, we discussed geographic hyperparameters that dictate how place is represented and the degree to which physical and human aspects of space are preserved. Choosing approaches to aggregation and similarity that do not preserve population density is a good first step towards addressing these challenges. For instance, researchers and engineers should consider adopting the S2-cell approach used by Weyand et al. [**321**], which are global, efficient, and naturally scale to reach a consistent number of data points across each cell, or census units that have been designed for the same task (e.g., census units, congressional districts). Instead of using physical distance, we have shown that measures like ordinal distance [**272**] can be more effective in measures of geographic similarity as they account for population density. Improving the quality of rural Wikipedia articles is a different challenge, but approaches that reduce the isolation of these articles and account for the shared history of these many towns—e.g., automatically detecting related sections in articles for entites that geographically contain a place—would be a large step towards reducing the higher per capita burden in rural areas. Similar approaches may help support languages with fewer speakers as well, making these solutions not just about rural regions.

A complementary theme arose in Part I that reflects an additional challenge to achieving representative coverage: *local content often provides greater value.* An important component of the success of online platforms like Wikipedia has been the ability for anyone to contribute content from or about anywhere [**136**]. We found, however, that for geographic user-generated content, this non-local content was insufficient in many contexts. This suggests that non-local contributions are not the full solution to achieving equal representation, and therefore we can hardly say that areas with large proportions of non-local content are

well-represented online. In social media (§3), we found that a substantial proportion of content was from non-local contributors in rural areas and that this content could alter the conclusions of research about these areas. In peer production (§4), we found a strong reliance on non-local contributors and bots allowed many rural areas to achieve excellent coverage but that local content was generally of higher quality. The research in those chapters suggested that filtering out non-local content, while not always important in urban areas given the wealth of local content, can be a key step to better representation of these areas. Kariyaa et al. [**164**] delve more deeply into this theme, but it complements discussion of the value of localness as expertise in neogeography [**109**] and additional research such as the distribution of local sources in Wikipedia [**274**] and the value of local contributions on OpenStreetMap [**73**].

10.2. On Evaluating Geographic Algorithms

The second half of this thesis focused on evaluating geographic algorithms. Again, alongside specific findings—e.g., no evidence of neighborhood avoidance in Google Maps or Mapquest (§9)—a few higher-level contributions are discussed below.

10.2.1. Hypothesizing about provider-fairness is challenging in geographic contexts.

The studies of place recommendation (§8) and vehicle routing (§7) focused on provider-fairness—i.e. not just whether these systems are accurate and unbiased for their users but also what is their impact on the items that they are recommending, be they the restaurants being recommended or the neighborhoods through which drivers are being directed. In both cases, alternative approaches to the standard approach to these algorithms were tested in

order to understand how they might affect provider-fairness. Neither sets of results were straightforward. For place recommendation, transforming the individual input user-visit data to reduce locality bias—i.e. less strongly encode neighborhoods—actually resulted in a lower diversity of neighborhoods being recommended across the universe of users. For vehicle routing, optimizing for a concept like "safety", which correlated with avoiding low-income neighborhoods, or "beauty", which correlated with high-income neighborhoods, resulted in mixed effects across New York City and San Francisco regarding the median income of neighborhoods through which the routes traveled. The deviations in routes required to avoid certain areas or prefer others led to inconsistent overall effects even if there were shared characteristics associated with those areas.

Together, these two studies suggest that provider fairness does not have straightforward ties to the biases encoded in the underlying data or model. This has two implications: 1) it suggests that, much as approaches to fairness-based machine learning have been developed for concepts of consumer fairness, the same will have to be true for provider fairness, and, 2) it demonstrates the importance of having representative datasets of inputs into these economically and socially impactful algorithms, as robustly evaluating their fairness relies on a representative set of queries.

Further exploration of provider fairness is essential in other recommendation domains such as whether music recommendation technologies such as Spotify or Pandora are equitable to the artists, product recommendation technologies such as Amazon are equitable to the sellers, loan or crowdfunding technologies like Kiva and Kickstarter are equitable for different regions or projects, or article recommendation technologies on Wikipedia are equitable for different types of article categories. It is likely that each domain will require specific evaluations and interventions, but developing methods and a vocabulary to discuss

provider fairness will be key to furthering the discussion of who these technologies should be benefiting.

10.2.2. An algorithm or sociotechnical system that works one way in one area will not necessarily work the same way in another area.

As first discussed in Hecht and Terveen [**141**], we find that evaluations of the efficacy and fairness of geographic technologies reach different conclusions depending on the setting in which they are studied. In this dissertation, we saw this with urban and rural performance in geolocation algorithms (§5) and the relationship between alternative routing algorithms and wealth (§7). Other researchers have seen similar effects, as discussed in Section 2.5.4.

This means that there is no single "representative" region in which an algorithm can be evaluated. This conclusion raises a number of interesting questions, namely how to identify a tractable number of areas in which to study the effects of a given algorithm. It might be for a mapping platform like Waze, that making adjustments to their algorithm requires careful evaluation of all major cities and rural areas. It would be useful to determine, however, if there are regions that are prototypical of a particular context and should be carefully examined at a minimum. The field of geography, with its many subfields and long history of classifying regions, would be able to guide some of this development.

10.2.3. Auditing geographic algorithms should not be limited to the descriptive.

Geographic algorithms are very challenging to audit and propose alternative approaches. They often involve highly sensitive data such as directions or restaurant queries that are difficult to anonymize and can reveal private details of individuals' lives. As such, datasets of how these services are used are not largely available and the algorithms themselves are

generally black boxes. As discussed in the prior sections, these algorithms neither have consistent impacts that are independent of place nor straightforward relationships between their biases and fairness of recommendations—i.e. these fine-grained details about usage are incredibly important to evaluating the impact of these technologies.

In our work, we have provided a blueprint for this approaching this challenge: for each algorithm, we have implemented it via open-source code and developed a reasonable proxy for data. For instance, we developed a version of Waze's safety routing algorithm per descriptions made by them in the media and used taxi routes as a proxy for driving directions origins and destinations (§7). We implemented a common approach to place recommendation and used Yelp reviewers as a proxy for queries from Yelp users (§8). We implemented two geolocation inference algorithms (§5) and collected supporting Twitter datasets. Implementing the algorithm instead of just relying on APIs (though often we examine both in the research) has allowed us to explore how manipulating specific components of the algorithm directly changes the impact. While our implementations and data sources almost certainly differ in important ways from what is being developed by industry, our approach still provides important data points to fuel discussions about what these algorithms should be optimizing given that they often are components of very impactful but largely black-box platforms.

10.2.4. Geographic hyperparameters should not just be left to default choices.

This thesis contains extensive related work (§2.5) on the choices that are made in representing geographic place and similarity. I detail how these choices encode structural inequalities in social science research (§3), geolocation algorithms (§5), place recommendation (§8), and geographic representation learning (§9). I also discuss alternatives to these choices in terms

of the data used to define place, how distance is defined, and how the data is aggregated. This framework provides a method for carefully considering how geography is encoded within technologies.

The difficulty and ad-hoc nature of spatial analysis is a recurring theme that, while evident to the researcher, is not necessarily evident in the publication of research that entails this type of work. For projects in which point data—e.g., Wikipedia articles, geolocated tweets, Yelp reviews—is to be related to place data such as socioeconomic statistics, there are many choices and preprocessing steps to link these two data sources. These choices are often driven by data availability—e.g., urban-rural codes are readily available for US counties but not other units such as census tracts. The goal of developing this framework is enumerate the choices available and provide some guidance for making these decisions from the standpoint of fairness. I hope to be part of continued work to help communicate the effects of choosing different geographic hyperparameters, much as has been done with related concepts of model fairness.[1,2]

10.3. Future Directions

There is an unanswered question in this thesis about what approach to take when a geographic technology is determined to be unfair. Chapter 8 explores how place recommendation technologies might be redesigned with an eye towards locality bias (not so strongly encoding location) and fairness (a more equitable distribution of recommendations). While I do believe strongly that we need robust means of evaluating the fairness of these sorts of technologies, I do not necessarily believe that when inequalities arise in these technologies, we should rely on the designers of these algorithms to balance the outcomes. There will

[1]https://pair-code.github.io/what-if-tool/
[2]https://research.google.com/bigpicture/attacking-discrimination-in-ml/

likely be simple fixes that are worth implementing—e.g., any geographic hyperparameters that strongly encode population density, such as aggregating to counties, or location, such as a dependence solely on trace data, should likely be avoided or carefully structured and evaluated. The larger challenges around segregation and the degree to which geographic technologies reinforce these patterns is a broader societal discussion though.

There are many reasons to believe that a top-down algorithmic solution is not *the* solution. Approaches that exclude sensitive attributes that seek to debias the data either through removing sensitive attributes [**127**] or explicit debiasing [**76**] have generally be found to be incomplete, with the problematic information—e.g., gender bias—still being encoded in other aspects of the model and difficult to impossible to remove in full. Optimizing for certain definitions of fairness in machine learning can harm the very groups that were supposed to be protected [**53**]. In general, any definition of fairness imposes some cost in terms of accuracy or invalidation of another component of fairness [**54, 171**], meaning that the design of fair algorithms requires an explicit weighing of costs. Given these potential pitfalls and difficult decisions around which definition of fairness to optimize for, it is not clear that merely pushing designers of technology to explicitly consider fairness would lead to an improved societal outcome. Likewise, policy may be a potential solution, but it is not clear that ideas like those raised around regulating where Waze could direct vehicles [**208**] would clearly lead to a beneficial outcome. These technologies are complex and their effects vary by region as shown in Chapters 7 and 8.

An alternative approach is value-sensitive algorithmic design [**350**], which is a process for developing algorithms to more effectively match the needs of a given community through an iterative process that takes the community's feedback into consideration. This type of process is happening in systems like Wikipedia, where community-managed bots [**103**] are core to

its functioning and open APIs where users can easily define their own thresholds and costs around precision and recall [**120**] are available, open to discussion, and iteratively improved based on feedback. These are powerful examples of algorithms that have been developed openly and with priority given to feedback and flexibility to mean a given community's needs.

The solution may also not be algorithmic in many situations. A carefully-designed user interface can be very powerful for counteracting biases. NextDoor, a neighborhood-level social network, is a powerful example of this: they were seeing a large number of racially-biased police reports being filed through their app, but instead of seeking to filter these reports out algorithmically or through content moderation, they added friction to the submission process. This change in design successfully forced people to submit more carefully considered claims and resulted in a substantial drop in racist reports [**142**].

The intended goal of this work then is less about proposing a more fair design of geographic technologies and more about providing a framework for evaluating the impact of these technologies that can help guide the discussion about what they should be optimizing. The takeaways from these case studies in Part II suggest that reconsidering how we evaluate these technologies will not be straightforward. Much more research and collaboration with the communities affected by these technologies is needed before making decisions about how they should be designed, but I hope that the research contained within this dissertation demonstrates that the choice to retain the default approaches to technologies like place recommendation is one that continues to encode inequality to the detriment of communities already at a structural disadvantage.

10.4. Concluding Remarks

As stated in the beginning, this dissertation is motivated by concerns about the role of internet technologies in deepening societal inequality. Through three different studies across multiple platforms and regions, the first part provides a detailed accounting of how structural inequalities associated with the urban-rural divide have been preserved in the consumption of online content by users, research, and algorithms. We demonstrate that achieving equal participation on these platforms would be insufficient to achieve equal representation. Understanding that these systems preserve structural inequalities motivates the second part of this dissertation: in-depth evaluations of three major types of geographic algorithms, as a subset of technologies with widespread social and economic implications. We provide a framework for geographic hyperparameters that shows how structural inequalities arise in these technologies and what alternative choices might counteract these inequalities. We develop methods for evaluating the impact of several of these technologies from the standpoint of provider fairness—i.e. their impact on the communities that they represent.

References

[1] An Introduction to Yelp Metrics as of September 30, 2018, Sept. 2018.

[2] Aaronson, D., Hartley, D. A., and Mazumder, B. The Effects of the 1930s HOLC Redlining'Maps. *Working Paper*, No. 2017-12 (2017).

[3] Abbar, S., Mejova, Y., and Weber, I. You Tweet What You Eat: Studying Food Consumption Through Twitter. ACM Press.

[4] Abdullah, S., Murnane, E. L., Costa, J. M., and Choudhury, T. Collective Smile: Measuring Societal Happiness from Geolocated Images. In *CSCW* (2015), ACM Press.

[5] Agarwal, A., Beygelzimer, A., Dudk, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[6] Alivand, M., Hochmair, H., and Srinivasan, S. Analyzing how travelers choose scenic routes using route choice models. *Computers, Environment and Urban Systems 50* (Mar. 2015), 41–52.

[7] Amato, C., Amir, O., Bryson, J., Grosz, B., Indurkhya, B., Kiciman, E., Kido, T., Lawless, W. F., Liu, M., McDorman, B., Mead, R., Oliehoek, F. A., Specian, A., Stojanov, G., and Takadama, K. Reports of the AAAI 2016 Spring Symposium Series. *AI Magazine 37*, 4 (Jan. 2017), 83.

[8] An, C., and Rockmore, D. Improving Local Search with Open Geographic Data. ACM Press, pp. 635–640.

[9] Ananny, M., Karahalios, K., Sandvig, C., and Wilson, C. Auditing Algorithms from the Outside: Methods and Implications. In *ICWSM* (2015).

[10] Anselin, L. Under the hood Issues in the specification and interpretation of spatial regression models.

[11] Anselin, L. *Exploring Spatial Data with GeoDa: A Workbook*. Center for Spatially Integrated Social Science, 2005.

[12] Antin, J., Chi, E. H., Howison, J., Paul, S., Shaw, A., and Yew, J. Apples to oranges?: comparing across studies of open collaboration/peer production. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (2011), ACM, pp. 227–228.

[13] Backstrom, L., Sun, E., and Marlow, C. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW* (2010), ACM.

[14] Badger, E., Bui, Q., and Pearce, A. The Election Highlighted a Growing Rural-Urban Split. *The New York Times* (Nov. 2016).

[15] Ballatore, A., Bertolotto, M., and Wilson, D. C. The semantic similarity ensemble. *Journal of Spatial Information Science*, 7 (Dec. 2013).

[16] Bamler, R., and Mandt, S. Dynamic word embeddings. In *International Conference on Machine Learning* (2017), pp. 380–389.

[17] Bank, T. W. Individuals using the Internet (% of population). Tech. rep., 2018.

[18] Barocas, S. Allocative versus Representational Harms in Machine Learning, Mar. 2018.

[19] Barocas, S., and Selbst, A. D. Big data's disparate impact. *Cal. L. Rev. 104* (2016), 671.

[20] Bast, H., Delling, D., Goldberg, A., Mller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., and Werneck, R. F. Route planning in transportation networks. *arXiv preprint arXiv:1504.05140* (2015).

[21] Becker, H., Naaman, M., and Gravano, L. Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM* (2011).

[22] Beir, M. G., Panisson, A., Tizzoni, M., and Cattuto, C. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science 5*, 1 (Oct. 2016), 30.

[23] Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (Aug. 2013), 1798–1828.

[24] Bivand, R., Altman, M., Anselin, L., Assuno, R., and Berke, O. Package spdep.

[25] Bjelland, M. D., Montello, D. R., Fellmann, J. D., Getis, A., and Getis, J. *Human Geography: Landscapes of Human Activities*, 12 ed. McGraw-Hill, 2013.

[26] Bliss, L. A Mapping Machine Identifies Wealth From Space. *CityLab* (June 2017).

[27] Blodgett, S. L., Green, L., and O'Connor, B. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), pp. 1119–1130.

[28] Blumenstock, J., Cadamuro, G., and On, R. Predicting poverty and wealth from mobile phone metadata. *Science 350*, 6264 (2015), 1073–1076.

[29] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (2016), pp. 4349–4357.

[30] Boyd, D., and Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society 15*, 5 (2012), 662–679.

[31] Brock, A. From the blackhand side: Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media 56*, 4 (2012), 529–549.

[32] Brockmann, D., Hufnagel, L., and Geisel, T. The scaling laws of human travel. *Nature 439*, 7075 (2006), 462.

[33] Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis 28*, 4 (1996), 281–298.

[34] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (2018), pp. 77–91.

[35] Bureau, U. C. 2010 Census Urban and Rural Classification and Urban Area Criteria, Feb. 2015.

[36] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. Discriminating gender on Twitter. In *EMNLP* (2011), Association for Computational Linguistics.

[37] Burke, R., Sonboli, N., and Ordonez-Gauger, A. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency* (2018), pp. 202–214.

[38] Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.

[39] Center, P. R. Social Media Fact Sheet. Tech. rep., Feb. 2018.

[40] Cha, M., Gwon, Y., and Kung, H. T. Twitter Geolocation and Regional Classification via Sparse Coding. In *ICWSM* (2015).

[41] Chen, L., Mislove, A., and Wilson, C. Peeking Beneath the Hood of Uber. In *IMC* (2015), ACM Press.

[42] Chen, X., Zeng, Y., Cong, G., Qin, S., Xiang, Y., and Dai, Y. On Information Coverage for Location Category Based Point-of-interest Recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas, 2015), AAAI'15, AAAI Press, pp. 37–43.

[43] Cheng, C., Yang, H., King, I., and Lyu, M. R. Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, Ontario, Canada, 2012), AAAI'12, AAAI Press, pp. 17–23.

[44] Cheng, Z., Caverlee, J., and Lee, K. You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. *CIKM* (2010).

[45] Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. Exploring Millions of Footprints in Location Sharing Services. *ICWSM 2011* (2011).

[46] Chollet, F. Using pre-trained word embeddings in a Keras model, July 2016.

[47] Clewlow, R. R., and Mishra, G. S. Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States. Tech. Rep. UCD-ITS-RR-17-07, Institute of Transportation Studies, University of California, Davis, 2017.

[48] Colley, A., Wenig, N., Wenig, D., Hecht, B., Schning, J., Thebault-Spieker, J., Lin, A. Y., Degraen, D., Fischman, B., Hkkil, J., Kuehl, K., Nisi, V., and Nunes, N. J. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. ACM Press, pp. 1179–1192.

[49] Collins, J. L., and Wellman, B. Small town in the Internet society: Chapleau is no longer an island. *American Behavioral Scientist 53*, 9 (2010), 1344–1366.

[50] Compton, R., Jurgens, D., and Allen, D. Geotagging one hundred million twitter accounts with total variation minimization. In *IEEE BigData* (2014).

[51] Compton, R., Lee, C., Xu, J., Artieda-moncada, L., Lu, T.-c., Silva, L. D., and Macy, M. Using publicly visible social media to build detailed forecasts of civil unrest. *Security informatics 3*, 1 (2013), 1–11.

[52] Compton, R., Lee, C.-K., Lu, T.-C., de Silva, L., and Macy, M. Detecting future social unrest in unprocessed twitter data:emerging phenomena and big data. In *ISI* (2013), IEEE, pp. 56–60.

[53] Corbett-Davies, S., and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[54] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. ACM Press, pp. 797–806.

[55] Cranshaw, J., Hong, J. I., and Sadeh, N. The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City. *ICWSM* (2012), 58–65.

[56] Culotta, A. Estimating county health statistics with twitter. In *JSM Proceedings* (2014), ACM Press, pp. 1335–1344.

[57] Culotta, A. Reducing Sampling Bias in Social Media Data for County Health Inference. *JSM Proceedings* (2014).

[58] De Longueville, B., Smith, R. S., and Luraschi, G. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks* (2009), ACM, pp. 73–80.

[59] De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., and Lepri, B. The death and life of great Italian cities: a mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web* (2016), International World Wide Web Conferences Steering Committee, pp. 413–423.

[60] Delling, D., Goldberg, A. V., Goldszmidt, M., Krumm, J., Talwar, K., and Werneck, R. F. Navigation made personal: inferring driving preferences from GPS traces. ACM Press, pp. 1–9.

[61] DelReal, J. A., and Clement, S. Rural Divide. *The Washington Post* (June 2017).

[62] Diakopoulos, N., Friedler, S., Marcelo, A., Barocas, S., Hay, M., Howe, B., Jagadish, H., Unsworth, K., Sahuget, A., Venkatasubramanian, S., Wilson, C., Yu, C., and Zevenbergen, B. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.

[63] Diaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 412:1–412:14.

[64] Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE 6*, 12 (Dec. 2011).

[65] D'Onfro, J. Here are all the Google services with more than a billion users. *Business Insider* (Oct. 2015).

[66] Doshi, T. Introducing the Inclusive Images Competition, Sept. 2018.

[67] Dovi, S. Political Representation. *Stanford Encyclopedia of Philosophy* (Aug. 2018).

[68] Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop: HIAI* (2013).

[69] Dubey, A., Naik, N., Parikh, D., Raskar, R., and Hidalgo, C. A. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision* (2016), Springer, pp. 196–212.

[70] Duckham, M., and Kulik, L. Simplest Paths: Automated Route Selection for Navigation. In *International Conference on Spatial Information Theory* (2003), Springer, pp. 169–185.

[71] Durkin, E. Alexa's advice to 'kill your foster parents' fuels concern over Amazon Echo. *The Guardian* (Dec. 2018).

[72] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), ACM, pp. 214–226.

[73] Eckle, M. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. *Proceedings of the ISCRAM 2015 Conference* (2015). 00002.

[74] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. A latent variable model for geographic lexical variation. In *EMNLP '10* (2010), Association for Computational Linguistics.

[75] El Ali, A., Van Sas, S. N., and Nack, F. Photographer paths: sequence alignment of geotagged photos for exploration-based route planning. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), ACM, pp. 985–994.

[76] Elazar, Y., and Goldberg, Y. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 11–21.

[77] Elgin, B., and Robison, P. How Despots Use Twitter to Hunt Dissidents, Oct. 2016.

[78] Elsmore, S., Subastian, I. F., Salim, F. D., and Hamilton, M. VDIM: Vector-based Diffusion and Interpolation Matrix for Computing Region-based Crowdsourced Ratings: Towards Safe Route Selection for Human Navigation. In *MUM* (2014), MUM '14, pp. 212–215.

[79] Elwood, S., Goodchild, M. F., and Sui, D. Z. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers 102*, 3 (2012), 571–590. 00197.

[80] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway Feedback Loops in Predictive Policing. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA* (2018), pp. 160–171.

[81] Evangelho, J. 'Pokémon GO' Is About To Surpass Twitter In Daily Active Users On Android. *Forbes* (July 2016).

[82] Fast, E., Chen, B., and Bernstein, M. S. Empath: Understanding Topic Signals in Large-Scale Text. ACM Press, pp. 4647–4657.

[83] Feng, S., Cong, G., An, B., and Chee, Y. M. POI2vec: Geographical Latent Representation for Predicting Future Visitors. In *AAAI* (2017), pp. 102–108.

[84] Feng, S., Li, X., Zeng, Y., Cong, G., Chee, Y. M., and Yuan, Q. Personalized Ranking Metric Embedding for Next New POI Recommendation. In *IJCAI* (2015), pp. 2069–2075.

[85] Ference, G., Ye, M., and Lee, W.-C. Location Recommendation for Out-of-town Users in Location-based Social Networks. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management* (New York, NY, USA, 2013), CIKM '13, ACM, pp. 721–726.

[86] Fischer, E. Locals and Tourists.

[87] Fotheringham, A. S., and Wong, D. W. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A 23*, 7 (1991), 1025–1044.

[88] Foundation, W. Recognise the Internet as a human right, says Sir Tim Berners-Lee as he launches annual Web Index, Dec. 2014.

[89] Foundation, W. Technology doesnt have to increase inequality a message to the leaders at WEF, Jan. 2017.

[90] foursquare Blog. foursquare is joining the OpenStreetMap movement! Say hi to pretty new maps!, Feb. 2012.

[91] Fox, K. OpenStreetMap: 'It's the Wikipedia of maps'. *The Guardian* (Feb. 2012). 00000.

[92] Fu, K., Lu, Y.-C., and Lu, C.-T. TREADS: A Safe Route Recommender Using Social Media Mining and Text Summarization. In *SIGSPATIAL* (New York, NY, USA, 2014), SIGSPATIAL '14, ACM, pp. 557–560.

[93] Gabrilovich, E., and Markovitch, S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence* (2007), Morgan Kaufmann Publishers Inc., pp. 1606–1611.

[94] Gallardo, A., and Martin, N. Another Thing Disappearing From Rural America: Maternal Care. *ProPublica* (Sept. 2017).

[95] Gamino, J. U of C/Hyde Park history. *Chicago Maroon* (Oct. 2014).

[96] Gao, H., Tang, J., and Liu, H. Exploring Social-Historical Ties on Location-Based Social Networks. In *ICWSM* (2012).

[97] Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems* (Mar. 2014).

[98] Garcia-Gavilanes, R., Quercia, D., and Jaimes, A. Cultural dimensions in twitter: Time, individualism and power. *ICWSM 13* (2013).

[99] Gatersleben, B., Murtagh, N., and White, E. Hoody, goody or buddy? How travel mode affects social perceptions in urban neighbourhoods. *Transportation research part F: traffic psychology and behaviour 21* (2013), 219–230.

[100] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. Using deep learning and google street view to estimate the demographic makeup of the us. *arXiv preprint arXiv:1702.06683* (2017).

[101] Geiger, R. S., and Halfaker, A. Using Edit Sessions to Measure Participation in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2013), CSCW '13, ACM, pp. 861–870.

[102] Geiger, R. S., and Halfaker, A. When the Levee Breaks: Without Bots, What Happens to Wikipedias Quality Control. In *WikiSym + OpenSym 2013 conference proceedings* (2013).

[103] Geiger, R. S., and Halfaker, A. Open algorithmic systems: lessons on opening the black box from Wikipedia. *AoIR Selected Papers of Internet Research 6* (2016).

[104] Geisberger, R., Rice, M. N., Sanders, P., and Tsotras, V. J. Route planning with flexible edge restrictions. *Journal of Experimental Algorithmics 17*, 1 (July 2012), 1.1.

[105] Gilbert, E., Karahalios, K., and Sandvig, C. The Network in the Garden : An Empirical Analysis of Social Media in Rural Life. *CHI* (2008), 1603–1612.

[106] Glaeser, E. L., Kominers, S. D., Luca, M., and Naik, N. Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. *Economic Inquiry* (July 2016).

[107] Golledge, R. G. Defining the criteria used in path selection. *University of California Transportation Center* (1995).

[108] Gonzalez, R. An AI that Predicts a Neighborhood's Wealth from Space, June 2017.

[109] Goodchild, M. NeoGeography and the nature of geographic expertise. *Journal of location based services 3*, 2 (2009), 82–96.

[110] Goodchild, M. F. A geographer looks at spatial information theory. In *International Conference on Spatial Information Theory* (2001), Springer, pp. 1–13.

[111] Goodchild, M. F. Editorial: Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research 2* (2007), 24–32.

[112] Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. Uneven Geographies of User-Generated Information : Patterns of Increasing Informational Poverty Uneven Geographies of Knowledge. *Annals of the Association of American Geographers 287210*, 44 (2014).

[113] Graham, M., Straumann, R. K., and Hogan, B. Digital divisions of labor and informational magnetism: Mapping participation in Wikipedia. *Annals of the Association of American Geographers 105*, 6 (2015), 1158–1178.

[114] Graham, M., Zook, M., and Boulton, A. Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Institute of British Geographers 38*, 3 (2013), 464–479.

[115] Greenstein, S., and Zhu, F. Is Wikipedia Biased? *American Economic Review 102*, 3 (2012), 343–48.

[116] Hagerstrand, T., and others. Innovation diffusion as a spatial process. *Innovation diffusion as a spatial process.* (1968).

[117] Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design 37*, 1 (2010), 682–703.

[118] Haklay, M. Nobody wants to do council estates digital divide, spatial justice and outliers AAG 2012, Mar. 2012.

[119] Halfaker, A., Geiger, R. S., and Terveen, L. G. Snuggle: Designing for Efficient Socialization and Ideological Critique. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 311–320.

[120] Halfaker, A., Geiger, S. R., Morgan, J. T., Sarabadani, A., and Wight, A. ORES: Facilitating re-mediation of Wikipedias socio-technical problems.

[121] Halfaker, A., Kittur, A., Kraut, R., and Riedl, J. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (2009), ACM, p. 15.

[122] Halfaker, A., and Riedl, J. Bots and cyborgs: Wikipedia's immune system. *Computer 45*, 3 (2012), 79–82.

[123] Hall, A., McRoberts, S., Thebault-Spieker, J., Lin, Y., Sen, S., Hecht, B., and Terveen, L. Freedom versus Standardization: Structured Data Generation in a Peer Production Community. ACM Press, pp. 6352–6362.

[124] Hamilton, K., Karahalios, K., Sandvig, C., and Langbort, C. The image of the algorithmic city: a research approach. *Ann Arbor 1001* (2014), 48109.

[125] Han, B., Cook, P., and Baldwin, T. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* (2014), 451–500.

[126] Haque, S., Kulik, L., and Klippel, A. Algorithms for reliable navigation and wayfinding. In *International Conference on Spatial Cognition* (2006), Springer, pp. 308–326.

[127] Hardt, M. Equality of Opportunity in Machine Learning, Oct. 2016.

[128] Hardy, D., Frew, J., and Goodchild, M. F. Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science 26*, 7 (2012), 1191–1212.

[129] Hardy, J., and Lindtner, S. Constructing a desiring user: Discourse, rurality, and design in location-based social networks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017), ACM, pp. 13–25.

[130] Hargittai, E., and Litt, E. The tweet smell of celebrity success : Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society 13*, 5 (2011), 824–842.

[131] Harney, N. C., and Charlton, J. The Siege on South Peoria Street. *Chicago Reader* (Jan. 2000).

[132] Harvey, F. To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In *Crowdsourcing geographic knowledge*. Springer, 2013, pp. 31–42.

[133] He, J., Li, X., Liao, L., Song, D., and Cheung, W. K. Inferring a Personalized Next Point-of-Interest Recommendation Model with Latent Behavior Patterns. In *AAAI* (2016), pp. 137–143.

[134] He, X., He, Z., Du, X., and Chua, T.-S. Adversarial Personalized Ranking for Recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2018), SIGIR '18, ACM, pp. 355–364.

[135] Hecht, B., and Gergle, D. Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories. In *C&T* (State College, PA, 2009), pp. 11–19.

[136] Hecht, B., and Gergle, D. On the "localness" of user-generated content. *CSCW* (2010), 229.

[137] Hecht, B., and Gergle, D. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *CHI* (Atlanta, GA, 2010), CHI '10, ACM, pp. 291–300. ACM ID: 1753370.

[138] Hecht, B., Hong, L., Suh, B., and Chi, E. H. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI* (2011), ACM.

[139] Hecht, B., and Shekhar, S. Spatial Computing Massive Open Online Course (MOOC), 2014.

[140] Hecht, B., and Stephens, M. A Tale of Cities : Urban Biases in Volunteered Geographic Information. *ICWSM* (2014).

[141] Hecht, B., and Terveen, L. The Role of Human Geography in Collective Intelligence. In *Collective Intelligence* (2017).

[142] Hempel, J. For Nextdoor, Eliminating Racism Is No Quick Fix. *Wired* (Feb. 2017).

[143] Hendrix, S. Traffic-weary homeowners and Waze are at war, again. Guess whos winning?, June 2016.

[144] Hill, B. M., and Shaw, A. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one 8*, 6 (2013), e65782.

[145] Hirnschall, C., Singla, A., Tschiatschek, S., and Krause, A. Learning User Preferences to Incentivize Exploration in the Sharing Economy. *arXiv preprint arXiv:1711.08331* (2017).

[146] Hochman, N., and Schwartz, R. Visualizing instagram: Tracing cultural visual rhythms. In *ICWSM-SocMedVis* (2012), pp. 6–9.

[147] Hwang, M.-H., Wang, S., Cao, G., Padmanabhan, A., and Zhang, Z. Spatiotemporal transformation of social media geostreams: a case study of Twitter for flu risk analysis. ACM Press, pp. 12–21.

[148] Iacobacci, I., Pilehvar, M. T., and Navigli, R. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *ACL (1)* (2016).

[149] Ingold, D., and Soper, S. Amazon Doesnt Consider the Race of Its Customers. Should It? *Bloomberg* (Apr. 2016).

[150] Ingraham, N. Apple using TomTom and OpenStreetMap data in iOS 6 Maps app. *The Verge* (June 2012).

[151] Ingram, D., and Franco, S. 2013 NCHS urban-rural classification scheme for counties. *Vital Health Statistics 2*, 166 (2014).

[152] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science 353*, 6301 (2016), 790–794.

[153] Jeffries, S. How the web lost its way and its founding principles. *The Guardian* (Aug. 2014).

[154] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS) 25*, 2 (2007), 7.

[155] Johnson, I., and Hecht, B. Structural causes of bias in crowd-derived geographic information: Towards a holistic understanding.

[156] Johnson, I., Henderson, J., Perry, C., Schning, J., and Hecht, B. Beautiful but at What Cost?: An Examination of Externalities in Geographic Vehicle Routing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*, 2 (2017), 15.

[157] Johnson, I., Sengupta, S., Schning, J., and Hecht, B. The Geography and Importance of Localness in Geotagged Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2016).

[158] Johnson, M. Providing Gender-Specific Translations in Google Translate, Dec. 2018.

[159] Johnston, R., Manley, D., and Jones, K. Spatial Polarization of Presidential Voting in the United States, 19922012: The Big Sort Revisited. *Annals of the American Association of Geographers 106*, 5 (2016), 1047–1062.

[160] Jurgens, D. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM 13* (2013), 273–282.

[161] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM* (2015).

[162] Kafsi, M., Cramer, H., Thomee, B., and Shamma, D. A. Describing and Understanding Neighborhood Characteristics through Online Social Media. ACM Press, pp. 549–559.

[163] Kanevsky, D. Technology Change as the Great Equalizer, May 2012.

[164] Kariryaa, A., Johnson, I., Schning, J., and Hecht, B. Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 265.

[165] Karpathy, A., and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3128–3137.

[166] Kay, M., Matuszek, C., and Munson, S. A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. ACM Press, pp. 3819–3828.

[167] Keegan, B. C., and Brubaker, J. R. 'Is' to'Was': Coordination and Commemoration in Posthumous Activity on Wikipedia Biographies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), ACM, pp. 533–546.

[168] Kim, J., Cha, M., and Sandholm, T. SocRoutes: safe routes based on tweet sentiments. ACM Press, pp. 179–182.

[169] Kittur, A., and Kraut, R. E. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008), ACM, pp. 37–46.

[170] Klein, M., Zhao, J., Ni, J., Johnson, I., Hill, B. M., and Zhu, H. Quality Standards, Service Orientation, and Power in Airbnb and Couchsurfing. *PACM on Human-Computer Interaction 1*, 1 (2017).

[171] Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science* (2017).

[172] Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., and Mislove, A. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. ACM Press, pp. 121–127.

[173] Kline, R. R. *Consumers in the Country: Technology and Social Change in Rural America*. Johns Hopkins University Press, July 2002.

[174] Kochhar, R., and Fry, R. Wealth inequality has widened along racial, ethnic lines since end of Great Recession. Tech. rep., Pew Research Center, Dec. 2014.

[175] Kogan, M., Palen, L., and Anderson, K. M. Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy. In *CSCW* (2015), ACM Press, pp. 981–993.

[176] Kula, M. Metadata Embeddings for User and Item Cold-start Recommendations. In *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015.* (2015), T. Bogers and M. Koolen, Eds., vol. 1448 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 14–21.

[177] Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, P. K. Geographic Dissection of the Twitter Network. In *ICWSM* (2012).

[178] Kumar, V., Bakhshi, S., Kennedy, L., and Shamma, D. A. Modeling Characteristics of Location from User Photos. In *Proceedings of the 2017 ACM Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces* (New York, NY, USA, 2017), HUMANIZE '17, ACM, pp. 1–6.

[179] Kusmin, L. D. Rural America At A Glance: 2015 Edition. Tech. rep., United States Department of Agriculture, 2016.

[180] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems* (2017), pp. 4066–4076.

[181] Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. WP: clubhouse?: an exploration of Wikipedia's gender imbalance. In *WikiSym* (2011), ACM, pp. 1–10.

[182] Lampos, V., and Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (2010), IEEE, pp. 411–416.

[183] Landeiro, V., and Culotta, A. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).

[184] Le, Q., and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 1188–1196.

[185] Le Falher, G., Gionis, A., and Mathioudakis, M. Where Is the Soho of Rome? Measures and Algorithms for Finding Similar Neighborhoods in Cities. In *ICWSM* (2015).

[186] Letchner, J., Krumm, J., and Horvitz, E. Trip router with individualized preferences (TRIP): Incorporating personalization into route planning. In *AAAI* (2006), vol. 21, p. 1795.

[187] Li, L., Goodchild, M. F., and Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science 40*, 2 (Mar. 2013), 61–77.

[188] Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *SIGKDD* (2012), ACM.

[189] Li, S.-C., and Wu, Y. South Florida Internet Based Route Choice Survey, 2012.

[190] Li, T. J.-J., Sen, S., and Hecht, B. Leveraging advances in natural language processing to better understand Tobler's first law of geography. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2014), ACM, pp. 513–516.

[191] Li, X., Pham, T.-A. N., Cong, G., Yuan, Q., Li, X.-L., and Krishnaswamy, S. Where you instagram?: Associating your instagram photos with points of interest. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015), ACM, pp. 1231–1240.

[192] Li, Y., George, S., Apfelbeck, C., Hendawi, A. M., Hazel, D., Teredesai, A., and Ali, M. Routing Service with Real World Severe Weather. In *SIGSPATIAL* (New York, NY, USA, 2014), SIGSPATIAL '14, ACM, pp. 585–588.

[193] Lieberman, M. D., and Lin, J. You Are Where You Edit : Locating Wikipedia Contributors Through Edit Histories. *ICWSM* (2009), 106–113.

[194] Lih, A. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia*. Hyperion, Mar. 2009.

[195] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision* (2014), Springer, pp. 740–755.

[196] Lindamood, J., Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web* (2009), ACM, pp. 1145–1146.

[197] Liu, B., Fu, Y., Yao, Z., and Xiong, H. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), ACM, pp. 1043–1051.

[198] Luca, M. Reviews, reputation, and revenue: The case of Yelp. com. *Harvard Business School NOM Unit Working Paper No. 12-016* (2016).

[199] Maeda, T. N., Tsubouchi, K., and Toriumi, F. Next Place Prediction in Unfamiliar Places Considering Contextual Factors. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2017), SIGSPATIAL'17, ACM, pp. 76:1–76:4.

[200] Mahmud, J., Nichols, J., and Drews, C. Home Location Identification of Twitter Users. *ACM TIST 5*, 3 (July 2014), 1–21.

[201] Malik, M. M., Lamba, H., Nakos, C., and Pfeffer, J. Population Bias in Geotagged Tweets. In *ICWSM* (2015).

[202] Manley, E., Addison, J., and Cheng, T. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *Journal of Transport Geography 43* (Feb. 2015), 123–139.

[203] Manley, E., Cheng, T., and Emmonds, A. Understanding Route Choice using Agent-based Simulation. *Proceedings of the 11th international conference on GeoComputation* (2011).

[204] Manske, M. Reasonator - Meta, Sept. 2015.

[205] Markoff, J. Technology; Not a Great Equalizer After All? *The New York Times* (June 1999).

[206] Marshall, A. Crime Alerts Come to Brazilian Waze, Just in Time for the Olympics, July 2016.

[207] Mashhadi, A., Quattrone, G., and Capra, L. Putting ubiquitous crowdsourcing into context. *CSCW* (2013), 611.

[208] McCarty, M. The Road Less Traveled? Not Since Waze Came To Los Angeles, June 2016.

[209] McCorriston, J., Jurgens, D., and Ruths, D. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *ICWSM* (2015). 00000.

[210] McFarland, M. The case for almost never turning left while driving, Apr. 2014.

[211] McGee, J., Caverlee, J., and Cheng, Z. Location prediction in social media based on tie strength. In *CIKM* (2013), ACM Press, pp. 459–468.

[212] McGookin, D., Gkatzia, D., Hastie, H., and Edinburgh, U. K. Exploratory Navigation for Runners Through Geographic Area Classification with Crowd-Sourced Data. *MobileHCI* (2015).

[213] McIlwain, C. Racial formation, inequality and the political economy of web traffic. *Information, Communication & Society 20*, 7 (July 2017), 1073–1089.

[214] McMahon, C., Johnson, I. L., and Hecht, B. J. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies.

[215] Medelyan, O., Milne, D., Legg, C., and Witten, I. H. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies 67*, 9 (2009), 716–754.

[216] Meech, R. Open traffic data is the best traffic data, Dec. 2016.

[217] Menking, A., and Erickson, I. The Heart Work of Wikipedia: Gendered, Emotional Labor in the Worlds Largest Online Encyclopedia. In *CHI* (2015), ACM, pp. 207–210.

[218] Meta-Wiki. WikiWomen's Collaborative - Meta.

[219] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[220] Miller, H. J., and Goodchild, M. F. Data-driven geography. *GeoJournal 80*, 4 (Aug. 2015), 449–461.

[221] Milne, D., and Witten, I. H. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), ACM, pp. 509–518.

[222] Mislove, A., Lehmann, S., Ahn, Y.-y., Onnela, J.-p., and Rosenquist, J. N. Understanding the Demographics of Twitter Users. *ICWSM* (2011), 554–557.

[223] Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one 8*, 5 (Jan. 2013).

[224] Musthag, M., and Ganesan, D. Labor dynamics in a mobile micro-task market. In *CHI* (2013), ACM, pp. 641–650.

[225] Naaman, M., Zhang, A. X., Brody, S., and Lotan, G. On the Study of Diurnal Urban Routines on Twitter. *ICWSM* (2012), 258–265.

[226] Naik, N., Philipoom, J., Raskar, R., and Hidalgo, C. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2014), pp. 779–785.

[227] Najm, W. G., Smith, J. D., and Smith, D. L. Analysis of Crossing Path Crashes. Tech. Rep. DOT-VNTSC-NHTSA-01-03, National Highway Traffic Safety Administration.

[228] National Administration of Surveying, M. a. G. Surveying and Mapping Law of the Peoples Republic of China. Tech. rep., 2002.

[229] Newson, P., and Krumm, J. Hidden Markov map matching through noise and sparseness. In *SIGSPATIAL* (2009), ACM, pp. 336–343.

[230] Nguyen, D., and Eisenstein, J. A Kernel Independence Test for Geographical Language Variation. *Computational Linguistics* (2017).

[231] Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., and Konstan, J. A. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web* (2014), ACM, pp. 677–686.

[232] Noble, S. U. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

[233] Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. Mining user mobility features for next place prediction in location-based services. In *ICDM* (2012), IEEE, pp. 1038–1043.

[234] O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[235] Orban, E., McDonald, K., Sutcliffe, R., Hoffmann, B., Fuks, K. B., Dragano, N., Viehmann, A., Erbel, R., Jckel, K.-H., Pundt, N., and Moebus, S. Residential Road Traffic Noise and High Depressive Symptoms after Five Years of Follow-up: Results from the Heinz Nixdorf Recall Study. *Environmental Health Perspectives 124*, 5 (Nov. 2015).

[236] Patten, C. J., Kircher, A., stlund, J., Nilsson, L., and Svenson, O. Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention 38*, 5 (2006), 887–894.

[237] Pavalanathan, U., and Eisenstein, J. Confounds and Consequences in Geotagged Twitter Data. *EMNLP* (2015).

[238] Pennington, J., Socher, R., and Manning, C. D. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (2014), pp. 1532–1543.

[239] Perrin, A. Social Media Usage: 2005-2015. *Pew Research Center* (Oct. 2015).

[240] Perrin, A. Digital gap between rural and nonrural America persists. Tech. rep., Pew Research Center, May 2017.

[241] Pfeil, U., Zaphiris, P., and Ang, C. S. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication 12*, 1 (2006), 88–113.

[242] Piorkowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. *CRAWDAD dataset epfl/mobility (v. 2009-02-24)*. Feb. 2009. Published: Downloaded from http://crawdad.org/epfl/mobility/20090224.

[243] Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. Do all birds tweet the same?: characterizing twitter around the world. In *CIKM* (New York, NY, USA, 2011), CIKM '11, ACM, pp. 1025–1030. 00025.

[244] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. IP geolocation databases: Unreliable? *SIGCOMM 41*, 2 (2011).

[245] Poon, L. Waze Puts Safety Over Speed by Minimizing Left Turns, June 2016.

[246] Posti, M., Schning, J., and Hkkil, J. Unexpected journeys with the HOBBIT: the design and evaluation of an asocial hiking app. ACM Press, pp. 637–646.

[247] Priedhorsky, R., Culotta, A., and Y. Del Valle, S. Inferring the Origin Locations of Tweets with Quantitative Confidence. *CSCW 29* (2014), 997–1003.

[248] Pryzant, R., Chung, Y.-j., and Jurafsky, D. Predicting Sales from the Language of Product Descriptions.

[249] Quattrone, G., Capra, L., and De Meo, P. There's No Such Thing as the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In *CSCW* (2015), ACM Press, pp. 1021–1032.

[250] Quattrone, G., Mashhadi, A., Quercia, D., Smith-Clarke, C., and Capra, L. Modelling growth of urban crowd-sourced information. In *Proceedings of the 7th ACM international conference on Web search and data mining* (2014), ACM, pp. 563–572.

[251] Quercia, D., Schifanella, R., and Aiello, L. M. The shortest path to happiness: recommending beautiful, quiet, and happy routes in the city. In *HT* (2014), pp. 116–125.

[252] Quercia, D., Schifanella, R., Aiello, L. M., and McLean, K. Smelly maps: the digital life of urban smellscapes. *ICWSM* (2015).

[253] Raiman, J., and Raiman, O. DeepType: Multilingual Entity Linking by Neural Type System Evolution. *AAAI* (2018).

[254] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[255] Raymond, E. The cathedral and the bazaar. *Knowledge, Technology & Policy 12*, 3 (1999), 23–49.

[256] Reagle, J., and Rhue, L. Gender bias in Wikipedia and Britannica. *International Journal of Communication 5* (2011), 21.

[257] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (2009), AUAI Press, pp. 452–461.

[258] Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. Supervised text-based geolocation using language models on an adaptive grid. In *EMNLP-CoNLL* (2012), Association for Computational Linguistics.

[259] Rosenblat, A., and Hwang, T. Regional Diversity in Autonomy and Work: A Case Study from Uber and Lyft Drivers. *Data and Society Working Paper* (2016).

[260] Rosenblat, A., Levy, K. E., Barocas, S., and Hwang, T. Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination. *Policy & Internet 9*, 3 (2017), 256–279.

[261] Rossi, L., Williams, M. J., Stich, C., and Musolesi, M. Privacy and the city: User identification and location semantics in location-based social networks. *arXiv preprint arXiv:1503.06499* (2015).

[262] Rothstein, R. Race and public housing: revisiting the federal role. *Economic Policy Institute* (Dec. 2012).

[263] Rothstein, R. The racial achievement gap, segregated schools, and segregated neighborhoods: A constitutional insult. *Race and social problems 7*, 1 (2015), 21–30.

[264] Runge, N., Samsonov, P., Degraen, D., and Schning, J. No more Autobahn!: Scenic Route Generation Using Googles Street View. In *IUI* (2016), pp. 147–151.

[265] Ruths, D., and Pfeffer, J. Social media for large studies of behavior. *Science 346*, 6213 (Nov. 2014), 1063–1064.

[266] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW* (2010), ACM, pp. 851–860.

[267] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Can an Algorithm Be Unethical? *Ann Arbor 1001*, 48109, 1285.

[268] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination* (2014).

[269] Sang, J., Mei, T., Sun, J.-T., Xu, C., and Li, S. Probabilistic Sequential POIs Recommendation via Check-In Data. OCLC: 830028128.

[270] Santani, D., and Gatica-Perez, D. Loud and Trendy: Crowdsourcing Impressions of Social Ambiance in Popular Indoor Urban Places. In *Proceedings of the 23rd ACM International Conference on Multimedia* (New York, NY, USA, 2015), MM '15, ACM, pp. 211–220.

[271] Schwartz, R., and Halegoua, G. R. The spatial self: Location-based identity performance on social media. *New Media & Society* (2014), 1–18.

[272] Sen, S., Johnson, I., Harper, R., Mai, H., Olsen, S. H., Mathers, B., Vonessen, L. S., Wright, M., and Hecht, B. Towards domain-specific semantic relatedness: a case study from geography. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).

[273] Sen, S., Li, T. J.-J., Team, W., and Hecht, B. WikiBrain: Democratizing Computation on Wikipedia. In *OpenSym* (New York, NY, USA, 2014), OpenSym '14, ACM, pp. 27:1–27:10. 00000.

[274] Sen, S. W., Ford, H., Musicant, D. R., Graham, M., Keyes, O. S., and Hecht, B. Barriers to the Localness of Volunteered Geographic Information. In *CHI* (2015), ACM Press, pp. 197–206.

[275] Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., and Haklay, M. M. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science 31*, 1 (Jan. 2017), 139–167.

[276] Service, U. E. R. Population & Migration.

[277] Shah, S., Bao, F., Lu, C.-T., and Chen, I.-R. CROWDSAFE: Crowd Sourcing of Crime Incidents and Safe Routing on Mobile Devices. In *SIGSPATIAL* (2011), GIS '11, pp. 521–524.

[278] Shao, J., Kulik, L., and Tanin, E. Easiest-to-reach neighbor search. In *SIGSPATIAL* (2010), ACM, pp. 360–369.

[279] Shao, J., Kulik, L., Tanin, E., and Guo, L. Travel Distance Versus Navigation Complexity: A Study on Different Spatial Queries on Road Networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (New York, NY, USA, 2014), CIKM '14, ACM, pp. 1791–1794.

[280] Sharker, M. H., Karimi, H. A., and Zgibor, J. C. Health-optimal Routing in Pedestrian Navigation Services. In *SIGSPATIAL* (New York, NY, USA, 2012), HealthGIS '12, ACM, pp. 1–10.

[281] Shekhar, S., Feiner, S. K., and Aref, W. G. Spatial Computing. *Commun. ACM 59*, 1 (Dec. 2015), 72–81.

[282] Shirky, C. A Speculative Post on the Idea of Algorithmic Authority, Nov. 2009.

[283] Sicular, T., Ximing, Y., Gustafsson, B., and Shi, L. The urbanrural income gap and inequality in China. *Review of Income and Wealth 53*, 1 (2007), 93–126.

[284] Signorini, A., Segre, A. M., and Polgreen, P. M. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1n1 Pandemic. *PLoS ONE 6*, 5 (May 2011), e19467.

[285] Sigurbjrnsson, B., and Van Zwol, R. Flickr tag recommendation based on collective knowledge. In *WWW* (2008), ACM.

[286] Silver, N. The Most Diverse Cities Are Often The Most Segregated, May 2015.

[287] Silver, N. A Users Guide To FiveThirtyEights 2016 General Election Forecast, June 2016.

[288] Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR abs/1409.1556* (2014).

[289] Singhal, A. Introducing the Knowledge Graph: things, not strings, May 2012.

[290] Society, C. H. Expressways, 2005.

[291] Soeller, G., Karahalios, K., Sandvig, C., and Wilson, C. MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. ACM Press, pp. 867–878.

[292] Statt, N. Facebook is using billions of Instagram images to train artificial intelligence algorithms. *The Verge* (May 2018).

[293] Stephens, M. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal 78*, 6 (Dec. 2013), 981–996.

[294] Stuart, H. Waze Lets Israelis Avoid Palestinian Areas, but Not the Other Way Around, Oct. 2016.

[295] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology 59*, 6 (Apr. 2008), 983–1001.

[296] Suresh, H., and Guttag, J. V. A Framework for Understanding Unintended Consequences of Machine Learning. *AAAI* (2019).

[297] Sutton, P. C., Elvidge, C. D., Ghosh, T., and others. Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. *International Journal of Ecological Economics & Statistics 8*, S07 (2007), 5–21.

[298] Tapia, A. H., Lalone, N., MacDonald, E., Hall, M., Case, N., and Heavner, M. AURORASAURUS: Citizen Science, Early Warning Systems and Space Weather. In *HCOMP* (2014).

[299] Tashev, I. J., Couckuyt, J. D., Black, N. W., Krumm, J. C., Panabaker, R., and Seltzer, M. L. *Pedestrian route production*. Google Patents, Jan. 2012.

[300] Thai, J., Laurent-Brouty, N., and Bayen, A. M. Negative externalities of GPS-enabled routing applications: A game theoretical approach. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on* (2016), IEEE, pp. 595–601.

[301] Thebault-Spieker, J., Terveen, L., and Hecht, B. Towards a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction* (2017).

[302] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. YFCC100m: The New Data in Multimedia Research. *Commun. ACM 59*, 2 (Jan. 2016), 64–73.

[303] Tobler, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography 46* (June 1970), 234.

[304] Toyama, K. The Internet and Inequality. *Commun. ACM 59*, 4 (Mar. 2016), 28–30.

[305] Traunmueller, M., Fatah gen. Schieck, A., Schning, J., and Brumby, D. P. The Path is the Reward: Considering Social Networks to Contribute to the Pleasure of Urban Strolling. In *CHI EA* (New York, NY, USA, 2013), CHI EA '13, ACM, pp. 919–924.

[306] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing Depression from Twitter Activity. In *CHI* (2015), ACM Press, pp. 3187–3196.

[307] Tufekci, Z. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400* (2014).

[308] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM 10* (2010), 178–185.

[309] Venerandi, A., Quercia, D., and Saez-trumper, D. Measuring Urban Deprivation from User Generated Content. *CSCW* (2015).

[310] Vhaduri, S., Ali, A., Sharmin, M., Hovsepian, K., and Kumar, S. Estimating Drivers' Stress from GPS Traces. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2014), ACM, pp. 1–8.

[311] Vincent, N., Johnson, I., and Hecht, B. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale

Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 566.

[312] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv preprint arXiv:1501.06307* (2015).

[313] Wang, H., Terrovitis, M., and Mamoulis, N. Location Recommendation in Location-based Social Networks Using User Check-in Data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2013), SIGSPATIAL'13, ACM, pp. 374–383.

[314] Wang, T.-Y., Harper, F. M., and Hecht, B. Designing Better Location Fields in User Profiles. In *Proceedings of the 18th International Conference on Supporting Group Work* (New York, NY, USA, 2014), GROUP '14, ACM, pp. 73–80.

[315] Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S., and Zhou, X. Geo-SAGE: A geographical sparse additive generative model for spatial item recommendation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 1255–1264.

[316] Warncke-Wang, M., Cosley, D., and Riedl, J. Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration* (New York, NY, USA, 2013), WikiSym '13, ACM, pp. 8:1–8:10.

[317] Warncke-Wang, M., Ranjan, V., Terveen, L., and Hecht, B. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *ICWSM* (2015).

[318] Warncke-Wang, M., Uduwage, A., Dong, Z., and Riedl, J. In Search of the ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (New York, NY, USA, 2012), WikiSym '12, ACM, pp. 20:1–20:10.

[319] Weait, R. Wolfram Alpha is using OpenStreetMap data, June 2010.

[320] Weidmann, N. B., and Schutte, S. Using night light emissions for the prediction of local wealth. *Journal of Peace Research 54*, 2 (2017), 125–140.

[321] Weyand, T., Kostrikov, I., and Philbin, J. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision* (2016), Springer, pp. 37–55.

[322] Wikipedia, t. f. e. Template:Grading scheme, Sept. 2015.

[323] Wing, B. P., and Baldridge, J. Simple supervised document geolocation with geodesic grids. In *ACL* (2011), Association for Computational Linguistics.

[324] Wood, S. a., Guerry, A. D., Silver, J. M., and Lacayo, M. Using social media to quantify nature-based tourism and recreation. *Scientific reports 3* (2013), 2976.

[325] Wulczyn, E., Thain, N., and Dixon, L. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web* (2017), International World Wide Web Conferences Steering Committee, pp. 1391–1399.

[326] Xie, M., Yin, H., Wang, H., Xu, F., Chen, W., and Wang, S. Learning Graph-based POI Embedding for Location-based Recommendation. ACM Press, pp. 15–24.

[327] Xu, T., Ma, Y., and Wang, Q. Cross-Urban Point-of-Interest Recommendation for Non-Natives:. *International Journal of Web Services Research 15*, 3 (July 2018), 82–102.

[328] Yan, B., Janowicz, K., Mai, G., and Gao, S. From ITDL to Place2vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2017), SIGSPATIAL'17, ACM, pp. 35:1–35:10.

[329] Yang, C., Bai, L., Zhang, C., Yuan, Q., and Han, J. Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), ACM, pp. 1245–1254.

[330] Yang, J., Morris, M. R., Teevan, J., Adamic, L. A., and Ackerman, M. S. Culture Matters: A Survey Study of Social Q&A Behavior. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011).

[331] Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (2011), ACM, pp. 325–334.

[332] Yin, H., Sun, Y., Cui, B., Hu, Z., and Chen, L. LCARS: a location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), ACM, pp. 221–229.

[333] Yu, Y., and Chen, X. A Survey of Point-of-Interest Recommendation in Location-Based Social Networks. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).

[334] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (2013), ACM, pp. 363–372.

[335] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. Who, Where, when and What: Discover Spatio-temporal Topics for Twitter Users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, ACM, pp. 605–613.

[336] Yuan, Q., Cong, G., and Sun, A. Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (New York, NY, USA, 2014), CIKM '14, ACM, pp. 659–668.

[337] Zeveloff, N. Israelis Sue Waze Navigation App for Creating Neighborhood Traffic Jam, Dec. 2016.

[338] Zhang, A. X., and Counts, S. Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage. In *CHI* (2015), ACM Press, pp. 2603–2612.

[339] Zhang, A. X., Culbertson, B., and Paritosh, P. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media* (2017), ICWSM '17.

[340] Zhang, A. X., Noulas, A., Scellato, S., and Mascolo, C. Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *Social Computing (SocialCom), 2013 International Conference on* (2013), IEEE, pp. 69–74.

[341] Zhang, C., and Wang, K. POI recommendation through cross-region collaborative filtering. *Knowledge and Information Systems 46*, 2 (2016), 369–387.

[342] Zhang, J., Kawasaki, H., and Kawai, Y. A Tourist Route Search System Based on Web Information and the Visibility of Scenic Sights. In *Proceedings of the 2008 Second International Symposium on Universal Communication* (Washington, DC, USA, 2008), ISUC '08, IEEE Computer Society, pp. 154–161.

[343] Zhang, J.-D., and Chow, C.-Y. GeoSoCa: Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations. In *Proceedings of the*

38th International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2015), SIGIR '15, ACM, pp. 443–452.

[344] Zhang, Y., Li, B., and Hong, J. Understanding User Economic Behavior in the City Using Large-scale Geotagged and Crowdsourced Data. In *Proceedings of the 25th International Conference on World Wide Web* (2016), International World Wide Web Conferences Steering Committee, pp. 205–214.

[345] Zhao, K., Cong, G., Yuan, Q., and Zhu, K. Q. SAR: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on* (2015), IEEE, pp. 675–686.

[346] Zheng, D., Hu, T., You, Q., Kautz, H., and Luo, J. Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User's Online Photo Collections. In *ICWSM* (2015).

[347] Zheng, Y.-T., Yan, S., Zha, Z.-J., Li, Y., Zhou, X., Chua, T.-S., and Jain, R. GPSView: A Scenic Driving Route Planner. *ACM Trans. Multimedia Comput. Commun. Appl. 9*, 1 (Feb. 2013), 3:1–3:18.

[348] Zheng, Y.-T., Zha, Z.-J., and Chua, T.-S. Mining Travel Patterns from Geotagged Photos. *ACM TIST 3*, 3 (May 2012), 1–18.

[349] Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and Xie, X. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. ACM Press, pp. 295–304.

[350] Zhu, H., Yu, B., Halfaker, A., and Terveen, L. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 194.

[351] Zickuhr, K., and Smith, A. "28% of American Adults Use Mobile and Social Location-Based Services". *Pew Internet and American Life Project* (2011).

[352] Ziebart, B. D., Maas, A. L., Dey, A. K., and Bagnell, J. A. Navigate Like a Cabbie: Probabilistic Reasoning from Observed Context-aware Behavior. In *UbiComp* (New York, NY, USA, 2008), UbiComp '08, ACM, pp. 322–331.

[353] Zielstra, D., and Hochmair, H. Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2299 (2012), 41–47.

[354] Zielstra, D., Hochmair, H., Neis, P., and Tonini, F. Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap. *ISPRS International Journal of Geo-Information 3*, 4 (Nov. 2014), 1211–1233.

[355] Zielstra, D., Hochmair, H. H., and Neis, P. Assessing the Effect of Data Imports on the Completeness of OpenStreetMapAU nited S tates Case Study. *Transactions in GIS 17*, 3 (2013), 315–334.

[356] Zielstra, D., and Zipf, A. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE international conference on geographic information science* (2010), vol. 2010.

[357] ehek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 2010), ELRA, pp. 45–50.