Human vs. Machine: A Systematic Comparison of Village-Level Geocoding Precision

Inken von Borzyskowski* Patrick M. Kuhn†

April 29, 2018

Abstract

Geocoded data has become increasingly popular in political science research, with more than a dozen articles featured in the top journals in the last two years alone. While geocoded data opens up new avenues of inquiry and increases our ability to empirically assess theoretical predictions, much depends on the reliability of geocoding. When geocoding the location of survey respondents, towns, or projects, researchers can employ machine or hand coding. While machine coding is fast, transparent, and replicable, its reliability is questionable, especially when geocoding data from non-English speaking developing countries that lack reliable maps. We investigate the reliability of machine-coded geographic referencing tools by comparing them to hand-coded datasets (based on Afrobarometer data). Relying on various statistical validation techniques, we show that hand coding in sub-Saharan Africa outperforms a common machine coding program in terms of precision and quality. We also replicate a recently published study with machine-coding. Our findings add a cautionary note on using machine-coded geo-referenced data and the new wave of research relying on such measures.

*Assistant Professor, Department of Political Science, Florida State University; Email: i.Borzyskowski@fsu.edu. Corresponding author.

†Assistant Professor of Comparative Politics, School of Government and International Affairs (SGIA), University of Durham; E-mail: p.m.kuhn@durham.ac.uk.

Space and geography are important concepts in political science, as well as in related disciplines such as economics and geography. With the advent of affordable and easy to use geographic information systems (GIS) software, researchers have increasingly modeled space explicitly (e.g., Barkan, Densham and Rushton, 2006; Ichino and Nathan, 2013; Nemeth, Mauslein and Stapley, 2014; Bunte and Vinson, 2015; Warren, 2015). Geocoded data has become increasingly popular in political science research: a review of the top five political science journals and a conflict journal[1] between 2000 and 2017 reveals that the number of articles using geocoded data increased from just one article in 2001-2005, to 9 articles in 2006-2010, 27 articles in 2011-2015, and another 12 articles in 2016-2017 alone.

While geocoded data opens up new avenues of inquiry and increases our ability to empirically assess theoretical predictions, much depends on the reliability of geocoding. When geocoding locations researchers can employ machine or hand coding. While machine coding is cheap, fast, transparent, and replicable, its reliability is questionable, especially when geocoding data from non-English speaking developing countries that lack reliable maps. While machine geocoded datasets in developed countries with good underlying maps have a high quality, this is unlikely to be the case for developing countries. Despite its increased use, we know surprisingly little about the quality of machine-geocoded datasets. To date, there has not been a systematic analysis comparing machine to human geocoded data. How precise is geo-coding? Is it significantly different from human coding, and if so, what drives these differences? Do differences matter for substantive results, or under what conditions can machine coded data be relied upon?

This article addresses these questions by providing a systematic comparison between the two coding approaches for survey data from 20 sub-Saharan African countries. Given the countries' low level of development and the variety of official and native languages,

[1]Journals included in review are the American Political Science Review, American Journal of Political Science, Journal of Politics, International Organization, British Political Science Review, and the Journal of Peace Research

this can arguably be seen as a hard test for machine coding. We investigate the reliability of machine-coded geographic referencing by comparing them to human coding (based on the AidData/Afrobarometer team). Comparing quality and precision between human and machine geocoding, we show that currently human coding seems to outperform machine coding in sub-Saharan Africa. Given that result, we consider the implications for applied work by replicating Nunn (2010) using both human and machine coded data.

Our findings add a cautionary note on using machine-coded geo-referenced data, especially in the context of developing countries. First, there are significant differences in precision and quality between machine and human-coded geographic data, with human coding generally out-perfoming automated data generation. Second, these significant differences between human and machine codings vary considerably between countries and are systematically related to contextual factors, such as the level of urbanization, infrastructure, and service provision. In addition to more densely populated areas, those locations with paved roads and post offices are much more likely to be correctly geo-coded by a machine than locations without such characteristics. Since locations that cannot be geocoded are at times dropped from the estimation, this has the potential to generate significant bias in estimation results.

Our replication suggests that the conditions under whether machine/human coding makes a difference depends on whether the key variables are affected by this lower-quality coding. In our particular replication, the main results are robust as the effect of the lower-quality machine coding on the key explanatory variables is relatively small, and urban/infrastructure/ services are not key confounders in the regression model. We conclude that geocoding *ex post* is challenging, time consuming, and expensive, and therefore best done *ex ante* in the process of local data collection (e.g. fielding a survey). If researches do rely on *ex post* machine coding of geographic information, great care is required. Researchers should question data quality, especially in developing countries, and be aware that (1) the precision and the declared "quality" of machine geo-coding might be significantly worse than for human-coded

data, and (2) these differences are non-random, driven by the degree of urbanization and infrastructure, which are often key confounders. Researchers should not take data quality at face value and document that results are robust to potential biases of machine-geocoded data.

Setup

To investigate quality differences between human and machine geocoding, we machine-coded Afrobarometer (AB) round 4 data and compare it to AidData's human coding of that same survey round (BenYishay et al., 2017). Afrobarometer surveys are widely used in researching African politics and generally considered high quality. They cover a variety of countries varying in terms of development, colonial history, languages, and political regimes, making them an interesting test set.

In geocoding AB rounds 1-6, AidData followed a double-blind methodology, originally developed for geo-referencing development projects (Standow et al., 2011). Using a team of trained geocoders, coders used the hierarchy of place names provided by the survey to assign coordinates to each location. Geographic databases such as GeoNames, Google Maps, and OpenSteetMap were used by two independent coders to find coordinates, reviewing statoids, encyclopedias, Wikipedia, and government websites to confirm location hierarchy and type. When coders disagreed, the issue was arbitrated by a senior researcher. We take the resulting human-coded data as a benchmark against which to evaluate the machine coding.

The machine coding is done using the OpenCage Data's Geocoder (OCG) application programming interface (API) (Opencage Data Ltd., 2018) via the Stata module `opencagegeo` (Zeigermann, 2018).[2] OCG is built on opensource products and open data, including Nominatim and OpenStreetMap, GeoNames, Natural Earth Data, OpenGeoCode, Yahoo GeoPlanet, and Postcodes.io, and is widely used by both commercial and non-commercial enterprises. Unlike other geocoding APIs (e.g., from Google Maps, MapQuest, and Here Maps),

[2]Also available for `R`, see `https://cran.r-project.org/web/packages/opencage/opencage.pdf`.

the OCG API has several advantages: it has worldwide coverage, performs consistently, can parse the sparse location hierarchy provided by the AB, and does not restrict the use and storage of geocodes.

Before geocoding we pre-process the country, region, district, and town/village variables of the Afrobarometer to improve machine coding performance by trimming unnecessary white spaces, removing dashes and brackets, typing out abbreviations, replacing special characters, and making all letters lower case. AB 4 contains 4,294 unique town/village-district-region-country observations, which we submit to the OCG API for geocoding. In the following section we compare the geocoding using the OCG to the human geocoded AB round 4.

Comparing Human and Machine Geocoding

We begin by comparing coding precision. Figure 1 depicts the proportion of locations that were coded at the town/village, district, region, and country level, for both the Afrobarometer and the OpenCage Geocoder coordinates.

While machine coding seems to have been able to code a slightly larger proportion of locations at the town/village level (62% vs. 58%), it seems to perform worse overall than human coding. The human coded AB data was able to code all locations at a lower than country level, whereas the OCG was unable to code more than 13% of locations at any lower level than the country. Further, the AB coded more than 92% of all observations at either the district or town/village level, whereas the OCG managed to do so for only 72%. Moreover, looking more closely at the OCG coding suggests that OCG over-reports coding precision. Looking at the data we found several instances where a specific town/village could not be located in districts where the capital of said district had the same name as the district and where OCG coded the district capital and indicated that the observation was coded at the town/village level. Our overall quality assessment is further supported by the cross-tabulation reported in Table 1 below.

While there is a significant overlap with roughly 51% of locations located on the main

Figure 1: Comparison of Geocoding Quality

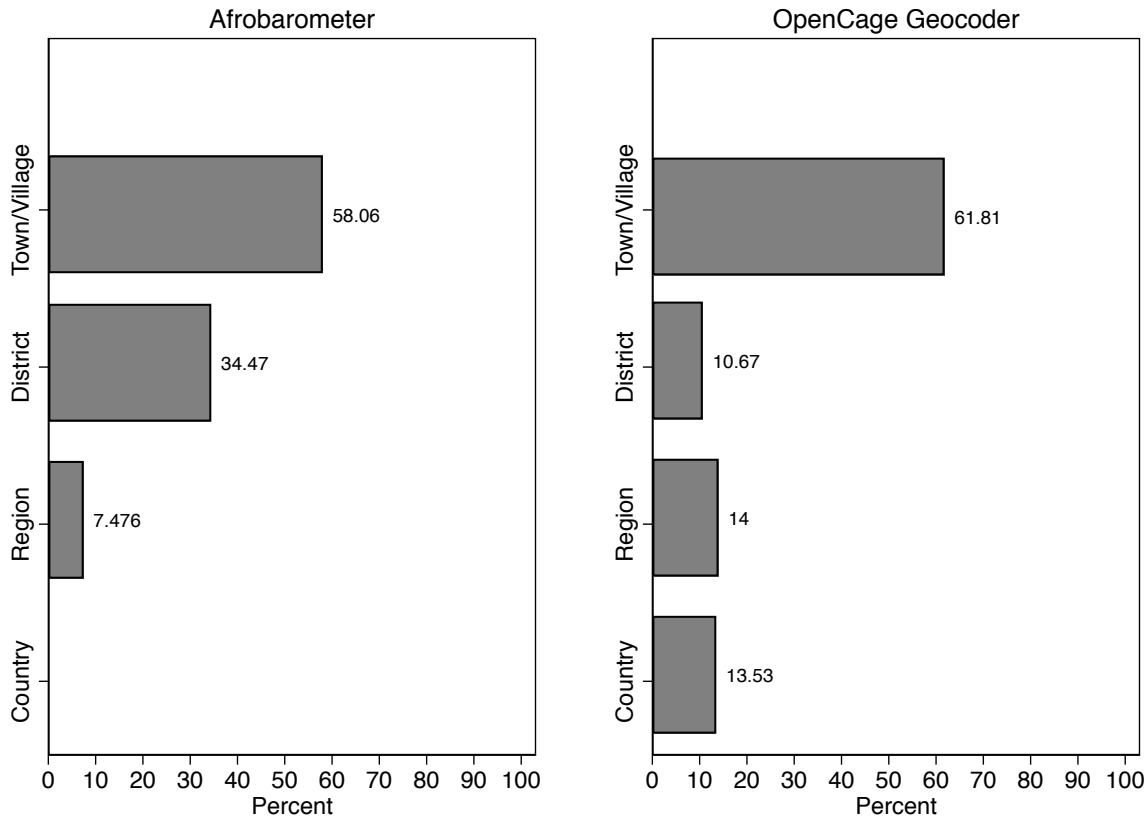


Table 1: Cross Tabulation of Geocoding Precision

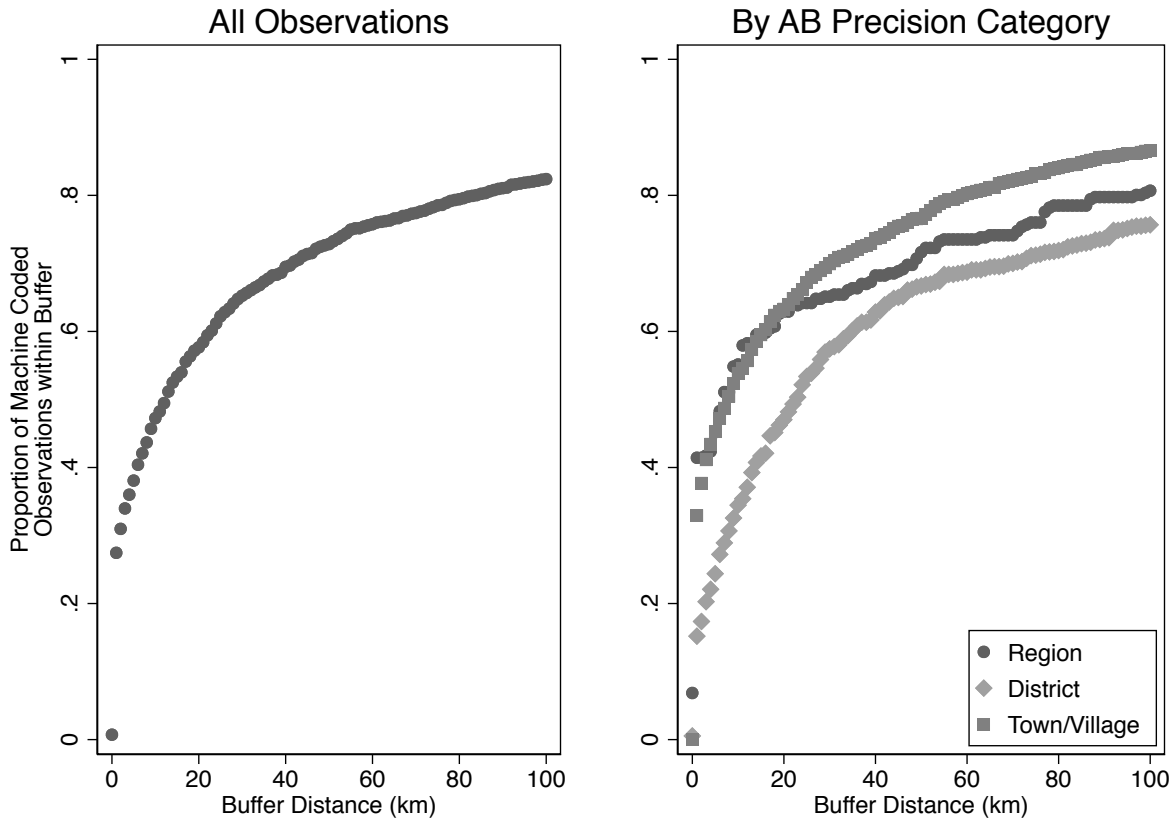
		OpenCage Geocoder				
		Country	Region	District	Town/Village	Total
Afrobarometer	Region	71	171	12	67	321
	District	331	101	240	808	1,480
	Town/Village	179	329	206	1,779	2,493
	Total	581	601	458	2,654	4,294

diagonal, there are fewer observations above the main diagonal (887) than there are below (1,217). This indicates that there are considerably more locations that were coded at a higher precision level by human compared to machine coding. Moreover, of the 887 observations above the diagonal, many of the 808 observations are cases where OCG could not identify the town/village, coded the district capital instead, and reported too high a precision level.

To compare geocoding quality beyond reported precision levels, we calculate the shortest

distance between the AB and OCG coordinates in kilometers (km). It ranges from 0km to 1,515km, with an average of 52.75km (see also Column 1 in Table 2) and a median of 12.45km, indicating that the distribution has a high positive skewness. To get a better sense of the distribution and differences across the AB precision categories, we calculate the proportion of machine coded observations that fall within differently sized buffers ranging from 0-100km (which is the range within which roughly 82% of all observations fall). Figure 2 shows those percentages for all observations (left) and split up by AB precision categories (right).

Figure 2: Distance Between Human and Machine Geocoding

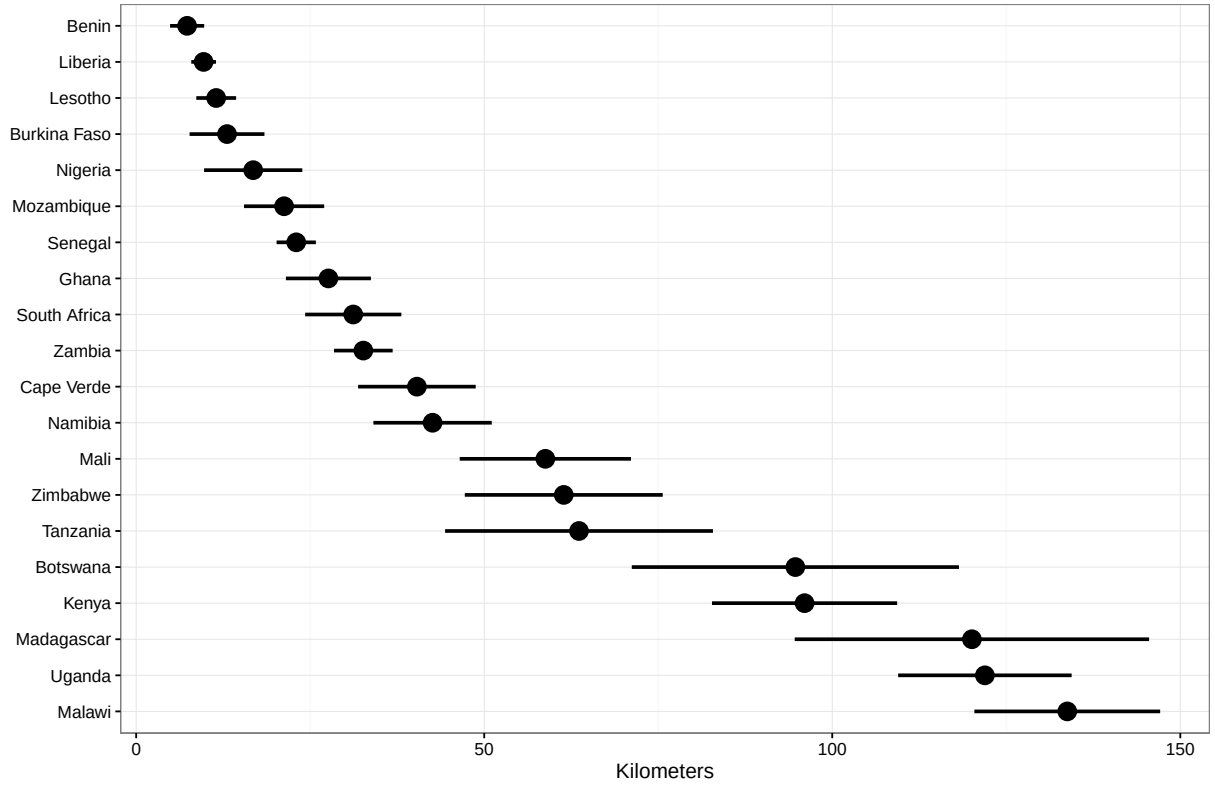


Looking at the left-hand graph, we see that only 0.7% of all observations match precisely, over 27% have a distance of less than 1km, and almost 70% are less than 40km apart. More interesting, however, is the right-hand graph, indicating that geocoding precision varies

across the AB precision categories. In line with our findings from Table 1, we observe that district level coded AB locations have the greatest average distance, followed by region, and town/village level coded locations. Once again, this difference is largely driven by OCG recording district capital codings wrongly at a too high precision level when the town/village could not be found.

Next we look the variation on average distance across the 20 sub-Saharan African countries included in the AB round 4. Figure 3 reports the average distance in km together with their 95% confidence interval.

Figure 3: Average Distance Between Human and Machine Geocoding



Note first, that all average distances are significantly greater than zero, suggesting that for none of the countries machine and human coding are indistinguishable. Next, note that the average distance varies quite a bit, ranging from 7km in Benin to over 133km in Malawi and that there are statistically and substantively significant differences between countries.

Additionally, note that there are no easily identifiable patterns with regard to country-level variables explaining these differences. South Africa, the wealthiest county in the list, is located in the middle, while Burkina Faso, one of the poorest countries listed, is ranked fourth. Moreover, countries with high (e.g., Benin, Ghana, and Botswana) and those with low democracy scores (e.g., Lesotho, Zimbabwe, Uganda) are both scattered throughout the ranking.

Finally, to investigate which location characteristics might be driving geocoding precision between human and machine coding, we ran a series of linear regressions with country-fixed effects on distance between AB and OCG. The results are reported in Table 2. The underlying survey questions and coding are detailed in the appendix.

Column 1 reveals the average distance across all 20 countries and that differences between countries explain roughly 20% of all variation. Column 2 includes a dichotomous variable for whether a location is urban or rural. Unsurprisingly, human and machine geocoding are significantly closer in urban than rural areas. Looking across the remaining two models (columns 3 and 4), the effect is consistently negative and large. On average the distance between the human and machine coded coordinates is 12.5km smaller for urban than rural locations. Column 3 adds two additive indices: one for infrastructure, ranging from 0 to 4, and one for services, ranging from 0 to 5. While the extent of services at a location does not seem to affect geocoding precision much, the extent of infrastructure has a statistically significant and large negative effect: human and machine coding are more similar for locations with extensive infrastructure compared to those with no infrastructure, even after controlling for whether a location is urban or rural. This finding is hardly surprising, as the provision of infrastructure often requires mapping, which feeds into geographic databases underlying machine coding. Lastly, Column 4 disaggregates the infrastructure and service indices to assess which components are driving the effect. Among the infrastructure components, paved roads and access to the sewage system seem to be most important. Being close to a paved road seems particularly important; it reduces average distance by almost as much

Table 2: Determinants of Distance Between Human and Machine Geocoding

	(1)	(2)	(3)	(4)
Constant	52.752***	59.920***	63.046***	59.764***
	(1.269)	(1.603)	(3.114)	(4.113)
Urban		-20.812***	-12.669***	-12.540***
		(2.657)	(3.579)	(3.593)
Infrastructure			-5.629***	
			(1.334)	
Electric Grid				4.872
				(4.160)
Pipped Water				-5.525
				(3.535)
Sewage Pipe Access				-7.598*
				(4.200)
Paved Road				-11.845***
				(3.269)
Services			1.241	
			(1.046)	
Post Office				-8.630**
				(4.173)
School				0.251
				(4.450)
Police Station				1.475
				(4.027)
Health Clinic				5.255
				(3.338)
Market				3.752
				(3.408)
Observations	4294	4294	4294	4294
R-Squared	0.205	0.215	0.218	0.222

Notes: Linear regression on distance between hand and machine coded coordinates in kilometers. All regressions include country fixed effects. Estimates significant at the 0.1 (0.05; 0.01) level are marked with * (**; ***). Robust standard errors.

as a location being urban rather than rural. Among the service components, having a post office seems important: human and machine coding are on average more than 8km closer for locations with a post office compared to those without.

Overall, our comparison suggests that there are significant differences between human and machine geocoded datasets in terms of accuracy. Human coding, at least in the case of sub-Saharan Africa is currently still more precise than machine coding. Machine coding quality varies drastically across countries and largely depends on the quality of underlying databases. Urban locations with greater infrastructure, especially paved roads, access to a sewage system, and a post office are more likely to be machine coded with high geographical precision.

Replication

Do the identified differences between human and machine coding matter substantively? To address this question, we replicate a published and widely cited study and compare its results using both human and machine coding. Under what conditions should geocoding differences matter for substantive findings? The extent to which geocoding differences matter for substantive results likely depends on how much these differences affect key variables in the model. If measurement differs significantly for outcome, explanatory, or key confounding variables, results are likely to differ as well. Conversely, if the key variables are largely unaffected by geocoding differences, then results are likely similar across human and machine coding.

We replicate a recent study by Nunn (2010) who uses human-coded geographic information to show that colonial Christian missions in Africa were effective in converting locals, and that these effects last until today. We choose this study because it uses geo-referenced data on the ethnic group and town/village-level, both as explanatory variable and key confounders, suggesting that geocoding precision might matter. The explanatory variable – number of historical mission stations near the respondent's town – may be influenced by

geocoding as it relies on the geocoding precision of respondents' town/village, which also matters for the constructing of various confounders and whether the location is urban or rural is controlled for.

Nunn (2010, 147) argues that today's descendants of ethnic groups which historically experienced greater missionary contact are today more likely to self-identify as Christian. To empirically assess this argument, Nunn links information on the geographic location of Christian missions in colonial Africa to geo-coded AB round 3 data and various location- and ethnic-specific controls. We attempt a scientific replication of the paper's main finding using AB data from round 4, which includes two more countries and more ethnic groups compared to AB round 3. Given that Nunn's argument is not specific to any time period or survey wave, our replication reveals to what extent his result is due to idiosyncrasies of AB round 3 data. In addition, we compare replication results using human and machine coding.

We construct our replication dataset by following the description in Nunn 2010, using data on various components (e.g., mission location, colonial rail road, and explorer routes). We use the Murdock 1959 data on pre-colonial ethnic groups (which Nunn provides on his website) and the human geocoded AB round 4 data, which we also machine coded as described above. A detailed description of variables included and necessary but minor coding adjustments between AB round 3 and 4 can be found in the appendix.

Using Nunn's (2010) model specification, we focus on the main effect (his Table 1, Columns 1-3), which is regressing self-reported Christian religion (a proxy for conversion) on ethnicity- and village-specific mission exposure, using logit models with country fixed effects, employment and living condition fixed effects, a range of individual-, ethnicity-, and village-level control variables, and clustered standard errors. Table 3 reprints the original study's results in Columns 1-3 and shows our results in Columns 4-6.

The scientific replication largely confirms the main finding from Nunn (2010). Qualitatively the results are similar: the coefficients on mission station in ethnic group and village are positive throughout and significant in most models. A closer look, however, does reveal a

Table 3: Scientific Replication of Table 1, Columns 1-3 in Nunn (2010)

	AB Round 3 (Nunn 2010)			AB Round 4 (Human Coded)		
	(1)	(2)	(3)	(4)	(5)	(6)
Missions stations among ethnic group	0.036***		0.144***	0.029		0.024***
	(0.011)		(0.044)	(0.028)		(0.006)
Missions stations in village		0.021***	0.033***		0.147***	0.086***
		(0.006)	(0.032)		(0.031)	(0.030)
Individual-level controls	✓	✓	✓	✓	✓	✓
Ethnicity-level controls	✓	✗	✓	✓	✗	✓
Village-level controls	✗	✓	✓	✗	✓	✓
Country fixed effects	✓	✓	✓	✓	✓	✓
Observations	20,755	20,775	20,775	22,538	22,538	22,538
Clusters	185	2,693	185/2,693	228	3,750	228/3,750
Pseudo R-Squared	0.28	0.28	0.29	0.33	0.32	0.33

Notes: The table reports logit estimates where the unit of observation is an individual. Coefficients are reported with (ethnicity/town/ethnicity-town) clustered standard errors in brackets. All regressions include country fixed effects. Individual-level controls include age, age squared, a gender indicator, five living condition fixed effects, six employment fixed effects, and an indicator for whether the respondent lives in an urban location. Ethnicity-level controls include an indicator variable that equals one if the ethnicity was contacted by a European explorer prior to the colonial period, an indicator variable that equals one if a railway line dissected the land inhabited by the ethnicity during the nineteenth century, a measure of the fraction of land suitable for cultivation and the fraction of land within ten kilometers of a water source, and the log normalized number of slaves exported during the Atlantic and Indian Ocean slave trades. The village-level controls include the same set of control variables but measured at the village level. Estimates significant at the 0.05 (0.01) level are marked with ** (***).

few differences. Our effect of the ethnic measure is estimated more noisily; the standard error in Column 4 is larger than Column 1, while the coefficient size is fairly similar. This might be due to issues of merging a larger number and more fine-grained ethnic groups reported in AB round 4 (see Appendix) or due to a weakening of the effect when using more recent survey data. Furthermore, comparing Columns 2 and 5, the village effect in the replication is significantly larger than in the original study. The z-score for the test of statistical difference between these coefficients is 3.99. This might be due to differences in geocoding precision between the original study and AB 4; both of which were human coded. Lastly, comparing Columns 3 and 6, which is the preferred specification in the original study, the coefficient on the ethnic measure is statistically significant but is six times smaller here than in the original study. The coefficients on the ethnic measure in Columns 3 and 6 are statistically different, with a z-score of 2.7. Again, this might be due to matching ethnic groups, due to sampling chance in the Afrobarometer survey, or because the effect is indeed substantively smaller in more recent survey data. Overall, while qualitatively similar, our replication suggests that

conversion rates might not be quite as high as suggested in the initial study.

In addition to this scientific replication with more recent data, we also replicate findings by comparing estimates from human versus machine coding. Table 4 Columns 1 and 2 reprint the results from Table 3 Columns 5 and 6. Table 4 Columns 3 and 4 show the identical estimations based on machine coded measures, replacing the village-level mission variable from the hand-coding with the same measure from the machine coding. The results are quite similar, with no statistical differences between coefficient estimates between human and machine coding.

Table 4: Comparing Human and Machine Geocoding

	Human Coding		Machine Coding	
	(1)	(2)	(3)	(4)
Missions stations among ethnic group		0.024***		0.025***
		(0.006)		(0.006)
Missions stations in village AB	0.147***	0.086***		
	(0.031)	(0.030)		
Missions stations in village OCG			0.150***	0.075**
			(0.034)	(0.030)
Individual-level controls	✓	✓	✓	✓
Ethnicity-level controls	✗	✓	✗	✓
Village-level controls	✓	✓	✓	✓
Country fixed effects	✓	✓	✓	✓
Observations	22,538	22,538	22,538	22,538
Clusters	3,750	228/3,750	3,750	228/3,750
Pseudo R-Squared	0.32	0.33	0.31	0.33

Notes: The table reports logit estimates where the unit of observation is an individual. Coefficients are reported with (ethnicity/town/ethnicity-town) clustered standard errors in brackets. All regressions include country fixed effects. Individual-level controls include age, age squared, a gender indicator, five living condition fixed effects, six employment fixed effects, and an indicator for whether the respondent lives in an urban location. Ethnicity-level controls include an indicator variable that equals one if the ethnicity was contacted by a European explorer prior to the colonial period, an indicator variable that equals one if a railway line dissected the land inhabited by the ethnicity during the nineteenth century, a measure of the fraction of land suitable for cultivation and the fraction of land within ten kilometers of a water source, and the log normalized number of slaves exported during the Atlantic and Indian Ocean slave trades. The village-level controls include the same set of control variables but measured at the village level. Estimates significant at the 0.05 (0.01) level are marked with ** (***).

There are two potential explanations for why results are surprisingly similar between human- and machine-coded data. One potential reason is that the machine-coded data are of high quality and very precise. However, we know from the comparisons in the previous section

that this is not the case. There are, in fact, large differences in geographic precision between human and machine-coded data, and these differences have clear drivers. A second potential reason is that the systematic difference between human- and machine-coding do not result in systematic measurement differences in key variables, so that the documented differences do not matter in this particular case. Specifically, the explanatory variable (missions) might be correlated highly between human and machine-coded measures, and the inclusion of the urban indicator from the survey data might also account for some of the systematic measurement error. Indeed, the human- and machine-coded number of missions near a town/village are strongly correlated (r=0.73, p=0.000). Correlations are also high for the village-level control variables water (r=0.69), agricultural suitability (r=0.80), contact with pre-colonial European explorers (r=0.61), and connection to colonial railways (r=0.79); all of which are highly statistically significant (p=0.000).

Further, the inclusion of *urban* indicator does not seem to affect the main result. To examine the importance of this control variable, we replicate our Table 4 (which is Nunn's Table 1 models 2-3) by (1) omitting urban as a control variable, (2) running models on the urban sub-sample, and (3) on rural sub-sample separately. Results are in Table 5 in the Appendix and show that the results are robust across specifications and that the coefficient estimates between human- and machine-coded data remain statistically indistinguishable. The conversion mechanism is not significantly different between rural and urban areas, so results do not differ with less precisely geocoded data. While human and machine geocoding differ significantly and systematically across contextual factors, these differences do not matter for the main result of this replication study.

Conclusion

Geo-coded data are increasingly popular but have some inherent risks with regard to reliability, and thus inference. These risks are magnified in non-English speaking developing countries that lack reliable maps, and when geocoding is done by machines instead of hu-

mans. We provide a systematic comparison between the two coding approaches for survey data from 20 sub-Saharan African countries.

We find that human coding outperforms machine coding in sub-Saharan Africa to date in terms of both quality and precision. Automation was unable to code more than 13% of locations at any lower level than the country. Furthermore, the AB coded more than 92% of all observations at either the district or town/village level, whereas the OCG managed to do so for only 72%. Automated geocoding is significantly less precise and reliable than human coding, and these differences are driven by local contextual factors, which include urbanization, infrastructure, service provision, as well as country context. In addition to more densely populated areas, those locations with paved roads and post offices are much more likely to be correctly geo-coded by a machine than locations without such characteristics. Since locations that cannot be geocoded are at times dropped from the estimation, this can potentially generate significant bias in result estimates. Our analyses suggest that the degree to which these quality differences matter for changing results depends on whether one of the key variables in the study is affected by geocoding.

These findings have important implications and add a cautionary note for research using machine-coded geo-referenced data. First, geocoding *ex post* is difficult. This applies to human coding as well but is more severe for machine coding. Even specially trained geocoders at AidData/Afrobarometer could only code the location of about 58% of towns/villages in sub-Saharan Africa. This is staggering and highlights the need to take great care when working with geo-referenced data from the developing world, particularly when relying on low levels of aggregation.

Second, it is better to geocode *ex ante*, during the process of local data collection. Spatial information is important and should be collected while fielding the survey or other data collection. Since many surveys in the developing world are collected on tablets and geocoding technology is readily available and cheap, it is relatively easy for enumerators to download GIS coordinates for each interview. In order to mitigate ethical issues about linking inter-

viewee coordinates to interviewee responses – which might enable identification of individual households and thus generate potential risks to respondents – survey administrators can add random noise to geo-coordinates. This is done, for example, by the Demographic and Health Surveys (DHS) Program, and protects individuals while indicating a geographic "zone" of a few kilometers for researchers to use for merging with location-specific covariates. As the coordinate displacement is random, it should not bias coefficient estimates.

Third, if researchers use or generate *ex post* machine coding of geographic information, they need to be careful. Researchers should be skeptical of and investigate data quality, especially when studying developing countries. They should be aware that (1) the precision and the declared "quality" (level of coding) from automation is significantly worse than for human-coded data, and that (2) these differences are driven by location-specific factors, such as urbanization, infrastructure, and access to certain services. Differences in data quality can be consequential when these location-specific factors play a significant role in the estimation. At a minimum, researchers should not take data quality at face value, and document that results are robust to potential biases of machine-geocoded data.

References

Barkan, Joel, Paul Densham and Gerard Rushton. 2006. "Space Matters: Designing Better Electoral Systems for Emerging Democracies." *American Journal of Political Science* 50:926–939.

BenYishay, A., R. Rotberg, J. Wells, Z. Lv, S. Goodman, L. Kovacevic and D. Runfola. 2017. "Geocoding Afrobarometer Rounds 1 - 6: Methodology & Data Quality." Available online at `http://geo.aiddata.org`.

Bunte, Jonas and Laura Thaut Vinson. 2015. "Local power-sharing institutions and inter-religious violence in Nigeria." *Journal of Peace Research* 53:49–65.

Deconick, Koen and Marijke Verpoorten. 2013. "Narrow and Scientific Replication of The Slave Trade and the Origins of Mistrust in Africa." *Journal of Applied Econometrics* 28:166–169.

Ichino, Nahomi and Noah L. Nathan. 2013. "Crossing the Line: Local Ethnic Geography and Voting in Ghana." *American Political Science Review* 107(02):344–361.

Murdock, George Peter. 1959. *Africa: Its peoples and their culture history.* New York: McGraw-Hill.

Nemeth, Stephen, Jacob Mauslein and Craig Stapley. 2014. "The Primacy of the Local: Identifying Terrorist Hot Spots Using Geographic Information Systems." *Journal of Politics* 101:3221–3252.

Nunn, Nathan. 2010. "Religious Conversion in Colonial Africa." *American Economic Review* 100:147–152.

Opencage Data Ltd. 2018. "OpenCage Geocoder." `https://geocoder.opencagedata.com/`.

Standow, Daniel, Michael Findley, Daniel Nielson and Josh Powell. 2011. "The UCDP-AidData Codebook on Geo-referencing Foreign Aid, Version 1.1." Uppsala Conflict Data Program, Uppsala, Sweden: Uppsala University.

Warren, Camber. 2015. "Explosive connections? Mass media, social media, and the geography of collective violence in African states." *Journal of Peace Research* 52:297–311.

Zeigermann, Lars. 2018. "Opencagegeo: Stata Module for Geocoding." `http://fmwww.bc.edu/repec/bocode/o/opencagegeo.pdf`.

Appendix

A Variables Used in Table 2

We use most of the location-specific enumerator coded variables from the Afrobarometer dataset round 4 to identify systematic determinants of machine geocoding precision. The included variables are detailed below (AB variable name in brackets). We code variables as missing if indeterminable or missing data.

- **Urban:** Dichotomous variable that is 1 if an urban primary sampling unit and 0 otherwise (URBRUR)
- **Electric Grid:** Is there an electric grid that most houses can access? 1=Yes; 0=No (EA_SVC_A)
- **Pipped Water:** Is there a piped water system that most houses can access? 1=Yes; 0=No (EA_SVC_B)
- **Sewage Pipe Access:** Is there a sewage system that most houses could access? 1=Yes; 0=No (EA_SVC_C)
- **Paved Road:** Think of your journey here: Was the road at the start point of the primary sampling unit /enumeration area paved/tarred/concrete? 1=Yes; 0=No (EA_Road)
- **Infrastructure:** Additive index of the four dichotomous variables above (i.e., electric grid, pipped water, sewage pipe access and paved road) ranging from 0 to 4.
- **Post Office:** Is there a post office present or within easy walking distance? 1=Yes; 0=No (EA_FAC_A)
- **School:** Is there a school present or within easy walking distance? 1=Yes; 0=No (EA_FAC_B)
- **Police Station:** Is there a police station present or within easy walking distance? 1=Yes; 0=No (EA_FAC_C)
- **Health Clinic:** Is there a health clinic present or within easy walking distance? 1=Yes; 0=No (EA_FAC_D)
- **Market:** Is there a market present or within easy walking distance? 1=Yes; 0=No (EA_FAC_E)
- **Services:** Additive index of the four dichotomous variables above (i.e., post office, school, police station, health clinic, market) ranging from 0 to 5.

The only locations-specific variables provided that were not included are cell phone service (too many missing values) and the two security related variables, asking enumerators whether police or military was present.

B Scientific Replication of Nunn (2010)

We follow Nunn (2010) in constructing the relevant variables. Below we detail the construction of each variable and note when it differs from the original study. Data for the dependent and individual-level control variables come from the Afrobarometer dataset round 4 and are coded missing in the case of don't know/refused (AB variable name in brackets). We link ethnic-level variables to the survey data by using information on each respondent's ethnic group (Q79) and a previous mapping between this AB round and the Murdock ethnic groups by Deconinick and Verpoorten (2013).[3]

Outcome Variable:

- **Protestant/Catholic Indicator**: Indicator variable that is 1 if a respondent self-identifies as Christian (Catholic or Protestant: mainstream, Evangelical, Pentecoastal), and 0 otherwise (Q79)

Explanatory Variables:

- **Mission Stations among Ethnic Group:** Number of Protestant and Catholic missions within the pre-colonial homeland of an ethnic group in Murdock (1959)
- **Mission Stations in Village:** Number of Protestant and Catholic missions within 25km of a town/village

Individual-level Control Variables:

- **Male**: Indicator variable that is 1 if a respondent is male, and 0 otherwise (Q101)
- **Age**: Respondent's age in years (Q1)
- **Age squared**: Respondent's age in years squared (Q1)
- **Employment**: Respondent's 6-category employment status (Q94); this replaces the occupation measure from the original study which was included in AB3 but not AB4
- **Living condition**: Respondent's view of their present living conditions (Q4B): (1) very bad, (2) fairly bad, (3) neither good nor bad, (4) fairly good, or (5) very good.
- **Urban**: Indicator variable that is 1 if the town/village is urban, and 0 otherwise (URBRUR)

Ethnic-level Control Variables:

- **Access to drinking water:** Fraction of ethnic homeland based on Murdock's (1959) mapping of African pre-colonial ethnic groups that is within 10km of a fresh water lake or major river

[3]Matching ethnic groups is considerably more involved in AB4 than AB3 because AB4 includes more than double the ethnic groups and only a portion of them overlap with Nunn's AB3 data. This is because AB4 includes additional countries and much more fine-grained identities of ethnic groups. See the readme file of Deconinick and Verpoorten (2013) for one account of this.

- **Abundance of fertile soil:** Fraction of ethnic homeland based on Murdock's (1959) mapping of African pre-colonial ethnic groups that is suitable for growing rain-fed crops with intermediate input according to the UN Food and Agriculture Organization (FAO)
- **Explorer route:** Indicator variable that is 1 if an ethnic group was contacted by pre-colonial European explorers
- **Colonial railway lines:** Indicator variable that is 1 if an ethnic group was connected to the colonial railway network
- **Slave trade:** Logged number of slaves exported from an ethnic group (normalized over their historic area) during the Atlantic and Indian ocean slave trade

Village-level Control Variables:

- **Access to drinking water:** Fraction of the 25km radius around a town/village that is within 10km of a fresh water lake or major river
- **Abundance of fertile soil:** Fraction of the 25km radius around a town/village that is suitable for growing rain-fed crops with intermediate input according to the UN Food and Agriculture Organization (FAO)
- **Explorer route:** Indicator variable that is 1 if the 25km radius around a town/village was contacted by pre-colonial European explorers
- **Colonial railway lines:** Indicator variable that is 1 if the 25km radius around a town/village was connected to the colonial railway network
- **Slave trade:** Logged number of slaves exported from an ethnic group during the Atlantic and Indian ocean slave trade within which the town/village is located

C Additional Results

Table 5: Replication using Human and Machine Geocoding and Urban Measures

	No Urban Control		Urban Subsample		Rural Subsample	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Table 1 Model 2 in Nunn (2010)						
Missions stations in village AB	0.157***		0.099**		0.215***	
	(0.030)		(0.042)		(0.045)	
Missions stations in village OCG		0.157***		0.126***		0.180***
		(0.033)		(0.045)		(0.045)
Observations	22538	22538	7862	7862	14676	14676
Clusters	3750	3750	1146	1146	2633	2633
Pseudo R2	0.315	0.313	0.303	0.302	0.329	0.326
Panel A: Table 1 Model 3 in Nunn (2010)						
Missions stations among ethnic group	0.024***	0.025***	0.025**	0.024**	0.021***	0.025***
	(0.006)	(0.006)	(0.010)	(0.010)	(0.008)	(0.008)
Missions stations in village AB	0.097***		0.047		0.150***	
	(0.029)		(0.040)		(0.044)	
Missions stations in village OCG		0.083***		0.060		0.098**
		(0.030)		(0.041)		(0.043)
Observations	22538	22538	7862	7862	14676	14676
Clusters	7193	7193	2914	2914	4310	4310
Pseudo R-Squared	0.328	0.328	0.326	0.326	0.340	0.339

Notes: The table reports logit estimates where the unit of observation is an individual. Coefficients are reported with (town/ethnicity-town) clustered standard errors in brackets. All regressions include country fixed effects. Individual-level controls include age, age squared, a gender indicator, five living condition fixed effects, and six employment fixed effects. Panel B includes ethnicity-level controls, which are an indicator variable that equals one if the ethnicity was contacted by a European explorer prior to the colonial period, an indicator variable that equals one if a railway line dissected the land inhabited by the ethnicity during the nineteenth century, a measure of the fraction of land suitable for cultivation and the fraction of land within ten kilometers of a water source, and the log normalized number of slaves exported during the Atlantic and Indian Ocean slave trades. Panels A and B include village-level controls, which are the same set of control variables but measured at the village level. Estimates significant at the 0.05 (0.01) level are marked with ** (***).