Online Appendix: Measuring the Landscape of Civil War (v1.0)

Dr. Rex W. Douglass
University of California, San Diego

Dr. Kristen A. Harkness
University of St. Andrews
kh81@st-andrews.ac.uk

December 2017

This online appendix accompanies the paper "Measuring the Landscape of Civil War: Evaluating Geographic Coding Decisions with Historic Data from the Mau Mau Rebellion." It is intended to provide greater technical detail than was within the scope of the original article. Please visit the online repository (available here[1]) for the most up to date version of this document as well as full technical replication material for the paper.

Contents

[1] https://github.com/rexdouglass/MeasuringLandscape

1 Introduction

The structure of this document is designed to mirror that of the original article. Additional detail for each step of the analysis can be found under the common section heading found in quotes.

2 "Best Practices and New Tools for Building Geographic Conflict Data"

2.1 "Which Coordinate Sources?"

2.1.1 Gazetteers

We employ a number of different gazeteer sources for this analysis including:

1. The 1964 Official Kenya Gazetteer (United States Board of Geographic Names, 1964)
2. The U.S. Board of Geographic Names' database of foreign geographic feature names or NGA (http://geonames.nga.mil/gns/html/namefiles.html)
3. The National GeospatialIntelligence Agency's GeoNet Names Server or GeoNames (http://download.geonames.org/export/dump;
4. Google Maps' search API;
5. Bing Maps' search API;
6. OpenStreetMap (Kenya extract retrieved from http://download.geofabrik.de/africa.html);
7. Global Administrative Areas database or GADM (http://www.gadm.org);
8. Getty Thesaurus of Geographic Names (http://www.getty.edu/research/tools/vocabularies/tgn/index.htı
9. International Livestock Research Institute (http://192.156.137.110/gis/search.asp?id=386);
10. Wikidata (https://www.wikidata.org).

2.1.2 Gazetteer Coverage

We subset each gazeteer to just the region of interest covering the conflict. As you can see, the spatial coverage varied wildly from source to source, with some like Geonames having very densive coverage of this area to others like Openstreetmap having relatively sparse coverage. Coverage for both Google and Bing map APIs is based on just the set of entities returned when searching for the toponyms in our events data, not their full coverage of this area. We further exclude entities that did not have a textual label associated with them (e.g. many of OpentStreetMap's features).

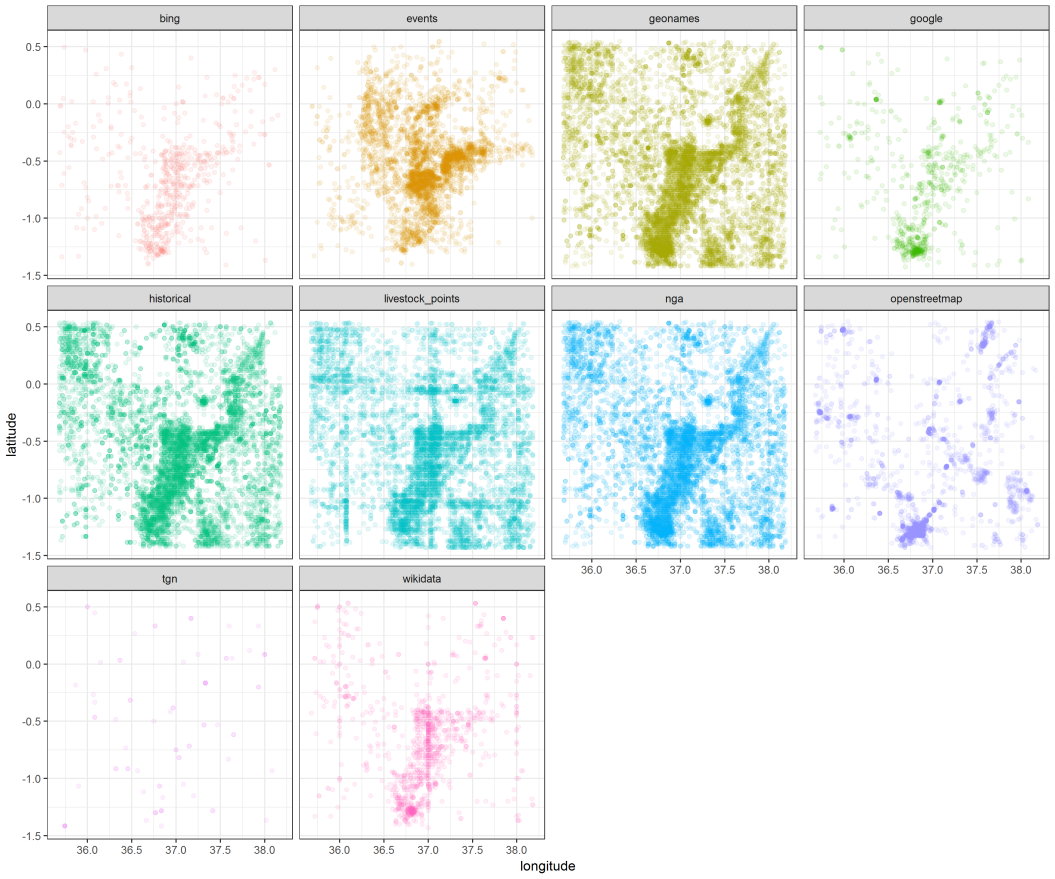

Figure 1: Spatial coverage within the region of interest of nine gazeteer sources that have point features.

2.1.3 Gazetteer Resolution

Each source has an implicit spatial resolution based on how it was created and whether and how those coordinates were translated (e.g. from degrees to decimal). By far the noisiest source was the digitization effort of the International Livestock Research Institute, with both coarse precision and noticeable artifacts of points lying directly on latitude and longitude map lines which were omitted or pushed to the side.[2] The historical gazetteer was presented in minutes and seconds, limiting its maximum precision to about 2km on the ground. NGA and geonames are a combination of the historical gazetteer as well as new points drawn from other sources with a plausibly higher resolution. Openstreetmap is drawn directly on satellite imagery by humans, but also imports some off the shelf data. Wikidata has a noticeable truncation to the whole degree. Our events locations have a known spatial resolution at the grid square, which our conversion to latitude/longitude gives a smooth but false sense of accuracy. In sum, when considering absolute maximum spatial precision we recommend a known error bounding box of at least 2km around any gazeteer entry.

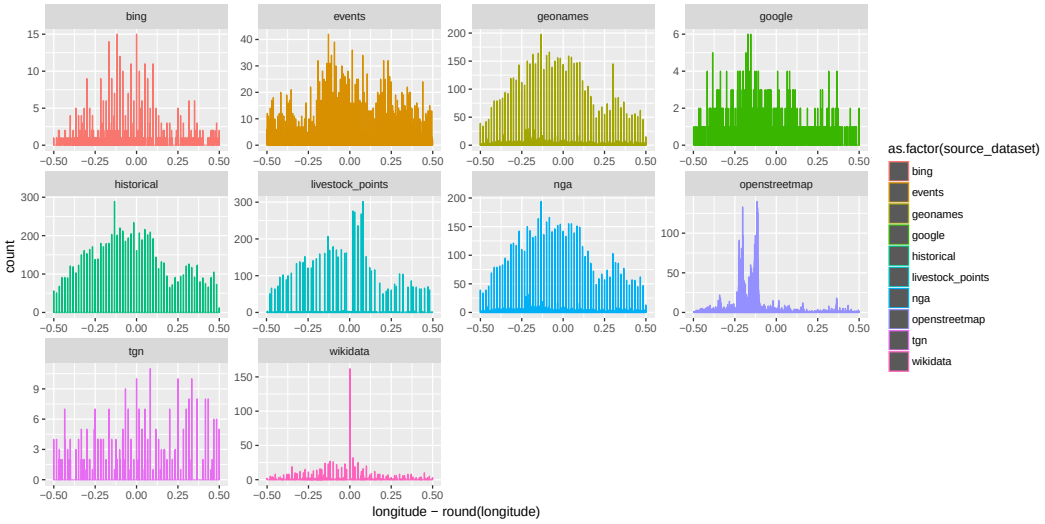

Figure 2: Histogram of longitude coordinates.

2.2 "Allow Self-Referential Matching?"

Building a gazeteer from source data itself is promising but potentially problematic. We wish to flag one concern in particular. If spatial auto-correlation is a concern, then self reference is a technique that builds in a mechanical spatial

[2] This appears to be an artifact of the automated process employed by a commercial vendor several decades ago.

correlation from points with shared names. For example, if the toponym was Nairobi, it represented a polygon covering a large area, but the only observed events with coordinates were in a single small ward of the city, then imputing those same points to other events that perhaps occurred in other wards, would mechanically introduce spatial auto-correlation as well as error.

2.3 "Which Geometry Type"

A number of our gazeteers included mixed geometry types, including points, polygons, and lines. How to deal with these correctly is an open research problem. We chose to take centroids, as points that were unbiasedly representative of the overall feature. In the future, researcher may prefer instead to sample from the entire feature in a way that is either more representative or reflects the uncertainty of representing a large area only a single point.

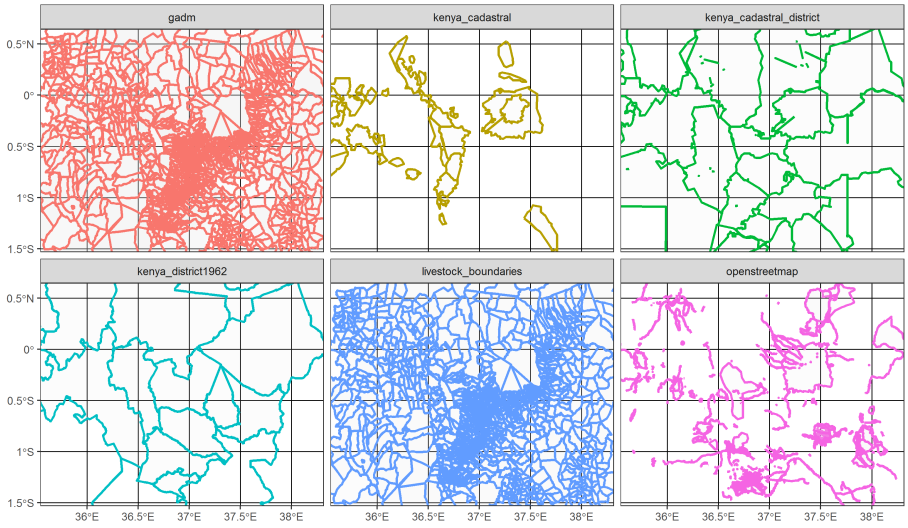

Figure 3: Histogram of longitude coordinates.

2.4 "Type of String Matching"

We introduce the concept of fuzzy matching but did not have space to describe the approach in the paper in detail. Our fuzzy matching pipeline has three major steps, stemming, locally sensitive hashing, and a supervised classifier.

2.4.1 The Problem

When comparing two toponyms, how should we decide when they are referring to the same real world location and when they are actually referring two different

real world locations? This is a difficult task for several reasons. The first is that relatively little information is conveyed by a string alone outside of the context in which it was used, e.g. the word "Paris" is not enough to uniquely distinguish between a vacation to Paris, France or a one to Paris, Texas.[3] Second, the way topynms are used (and abused) in natural language text leads to many different unique strings referring to the same underlying toponym. We have enumerated just some of the ways toponyms are misrepresented in natural language text in our event data, shown in tables below.

2.4.2 Training Data

We treat this a supervised learning problem and develop a hand labeled dataset of toponym matches (3,192) and mismatches (17,051). We construct a training dataset by taking events with both text and coordinates (5,128), pairing them with the ten gazetteer entries closest in geographic proximity. Removing duplicate diads reduces the number to 20,242 diads. Excluding exact matches reduces it further to 18,405 diads.

2.4.3 Stemming

Toponyms often include a mix of uniquely identifying tokens (stems) and generic place type descriptions (postfixes) like "farm", "estate", "hospital", etc. This creates a problem for string matching for two reasons. First, postfixes are not used consistently, appearing in some entries for the same place but not others. Second, postifixes make toponyms look similar because of their type but not necessarily their geography. We therefore develop a rule based toponym stemmer. We took the 14,885,300 unique names and alternate names, stripped off the first word of each, leaving 8,740,662 unique suffixes. To account for multiword stems and to ensure we capture broadly applicable postfixes we further require them to appear at least 5 times, resulting in 161,034 unique postfixes. For example, the five most common postfixes globally are "creek" (63,214), "lake" (40,595), "river" (36,379), "island" (31,347), "well" (30,091), "point" (28,225), and "cemetery" (27,345). Across the 57,903 unique toponyms/descriptions across both the events and combined gazeteers, we find 45,131 unique stems, and 2,546 unique postfixes.

2.4.4 Locally Sensitive Hashing

Unfortunately, the number of pairwise comparisons that must be made scales quadradically in the number of items, over a billion comparisons for our moderately sized dataset (k^2 / 2 = 45,131 stems^2 / 2= 1,018,403,580), and so using direct comparisons is impractical. We therefore introduce an intermediate culling stage based on approximate matching. The goal is to strike a balance that minimizes the number of false negatives at the cost of a reasonable number

[3] Having done both, an author would like to convey that there is a difference between these two vacations and that it is very important to know this beforehand.

Variant	Examples
Nonnames	
Personal Names	a.e.aggatt's farm; buckley's road; count van der stegen; lady eleanor cole's farm at elmentetia
Near Spellings Same	Burguret Estate;Burgaret Estate;Burgaret Entate
Near Spellings Different	s.e.bastard; S.S. Bastard's Farm
Directional	1 mile w.of bariaho; between eburru and kipipiri; between slaters and salisbury road; border of the iriaini & magutu, iriaini; embu/ south nyeri border; footpath leading to muthara; marula estate near longonot
Synonyms	aberdare forest; aberdare forest reserve
Supranational Information	aberdares f.hall; bentall's farm, nakuru; col.culton house kon-gaita
Abbreviations	aguthi loc s.n.r; clarke's fm.; gura rv.; malik s/mills; nr githumu
Corporate Names	blue post hotel
Ambiguous	
Spelling Errors	chehe sow mills; nairobi south esp llabour line
Contemporary Context	chief hinga's location; chief makimeis and waruiru's location; forest loc 14
Compound Locations	cole estates, gamble's farm
Grid Locations	eastings 43-47, northings 34-35
Sub-Information	eastleigh section 7
Compound Names	fort hall / thika road; gakurwa loc 14/15
Punctuation	luis farm; luis' farm
Numeric Postfix Irrelevant	marmanet mill no1
Numeric Postfix Relevant	Location 1; Location 11
Additional information	loc 8 fort hall
Directional Potfixes	mathira north; mathira south
Missing spaces	murandialocation 8

Table 1: Example natural language descriptions of toponyms.

	Events	gazeteers	News Corpus	Issues
"Ol Kalou"	"OLKALOU", "OL KALOU","Mark's farm OLKALOU","KALOU TOWN-SHIP","KALOU","01 KALOU TOWNSHIP","01 KALOU FLATS", and "01 KALOU"	"Ol Kalou" (PPLA, PPL, locality;political, PopulatedPlace, town, railway_station, Trading Centre) "Olkalou Country Club" (SOCF), "olkalou" (locality;political), "ol kalou flat" (locality;politica)," "kalou" (locality;political) "Ol Kalou Road" (unclassified)	kalou, o1 kalou, o kalou, o'kalou, ol kalou, oikalou, oi kalou, olkalou, okalou, kalou salient, kalou route, kalou hospital, kalou township, o'kalou plot, township of olkalou, kalou district, olkalou township, okalou plot, o1 kalou route, kalou west, kalou central, kalou nyandarua district, kalou road, oil kalou, oikalou hospital, kalou hospital nyandarua, kalou south, alou town, oj kalou, ol kalou township, township of olkalou township, kalou location, township of ol kalou, kalou naivasha district, o kalou salient, oi'kalou, ol kalou location	Consolidating different spellings and errors. Misspellings occur frequently in news corpus also. Levels of aggregation that appear in news corpus but not gazeteers (e.g. location, distric).
"BASTARD"	W.K. BASTARD'S PUMP HOUSE, W.K. BASTARD'S FARM, S.S.Bastard's Farm Nanyuki, S.S.Bastard's Farm,S.K.Bastard, Kongai R.W.K.Bastard, HUTCHINSON'S FARM, S.S BASTARDS FARM, BASTARD'S FARM	NONE!	bastard, s bastard, bastard esq, k bastard, w k bastard, k bastard esq, w k bastard esq, s s bastard, nanyuki bastard, bastard william, w s bastard, bastard esq po, h s bastard, s bastard esq, hs bastard, kenneth bastard, bastard s, william kenneth bastard, bastard esq farm, bastard esq nanyuki, bastard esq sw estate,	Zero gazeteer matches. Relying on coordinates from other events. Common surname Bastard applies to multiple households. Surname alone isn't sufficient to disambiguate which.
Nairobi	THIKA/NAIROBI ROAD,R NAIROBI GOLF CLUB , NAIROBI/NAKURU, NAIROBI WEST, Nairobi South E.A.P & L.Labour Line , NAIROBI SOUTH, NAIROBI RIVER, NAIROBI DAM, NAIROBI DAM, NAIROBI CITY, NAIROBI, KIAMBU, NAIROBI	464 matches... Nairobi, Nairobi City, Nairobi City County, Nairobi National Park, 2011 Nairobi pipeline fire, Bahati, Nairobi, Nairobi - Mombasa Railway, Nairobi - Nakuru Road, Nairobi / Dagoretti, NAIROBI AIRPORT		No events with Nairobi provide coordinates to verify strategy. Extreme number of gazetteer matches all over the country. Extreme number of text matches.

Table 2: Example toponym variations across sources.

of false positives. The approach we employ is called Locality Sensitive Hashing (LSH), and it proceeds in two steps (Leskovec et al. 2010, chapter 3). First, we convert each toponym into a standardized sparse vector of binary features. Q-gram distance performed well as a feature in later stages, so we choose 2-character ngram profiles as our sparse feature set.[4] The similarity between any two profiles is measured as the Jaccard distance which is the just the count of shared items in both sets divided by the total number of items in both. To avoid having to calculate that distance for every single pair of items, we apply a hash function that assigns two items to the same id with some probability as monotonic function of their Jaccard distance. By using multiple hashes, one can choose an arbitrary threshold for which items below a certain distance will be matched with very high likelihood. The advantage is this procedure only requires a single pass through the entire set of items. The downsides are that it is approximate, provides only a hard yes or no match, and there is a trade off between retrieving too many suggestions and losing too many real matches which must be hand tuned.

Using the hand labeled training data, we experimentally determined the optimal combination of gram features, and hash hyperparameters. We settled on 2-chaaracter ngrams, and LHS parameters of 100 minihashes, 25 bands, and 4 rows. On our training data, this produced a recall rate of 0.88 at the cost of an average of 29 suggestions per item. These pairs that pass the first sage's bar then go on to the second stage for further refinement and a direct estimate of the likelihood of a match.

2.4.5 Training and Test Split

For training the next stage of the classifier, we refine our dataset of hand labeled toponym and mismatches into a training and test split. First, to ensure the classifier is invariant to the ordering of the diad, we add the inverse of each diad and remove duplicates again, leaving 34,496 directed diads. To construct the test split, we first stem each toponym and then randomly select 500 stems to serve as a test split. Any diad, with a toponym on either end that shares one of those 500 stems is excluded for the testing split. The final division is 27,161 training directed diads and 5,521 testing directed diads, with 1,815 and 308 positive examples respectively.

2.4.6 Fuzzy Matcher

We train a classifier to predict the likelihood of match between two toponyms as a function of properties of each. We consider both the full strings and also their stems. We employ a number of string distance measures as features (Santos et al. 2017), described in the table below. For each, we employ a number of types of string distance (Jaro, Optimal String Alignment, Levenstein, Damerau

[4]Grams are simply counts of characters {'a','b','1'}, pairs of characters {'ed','ac'} , sets of three characters in a row {'abc','ing'}, etc. Skip grams extend this idea to allower for gaps between pairs or tripplets, e.g. {'a_c','c_e'}, etc.

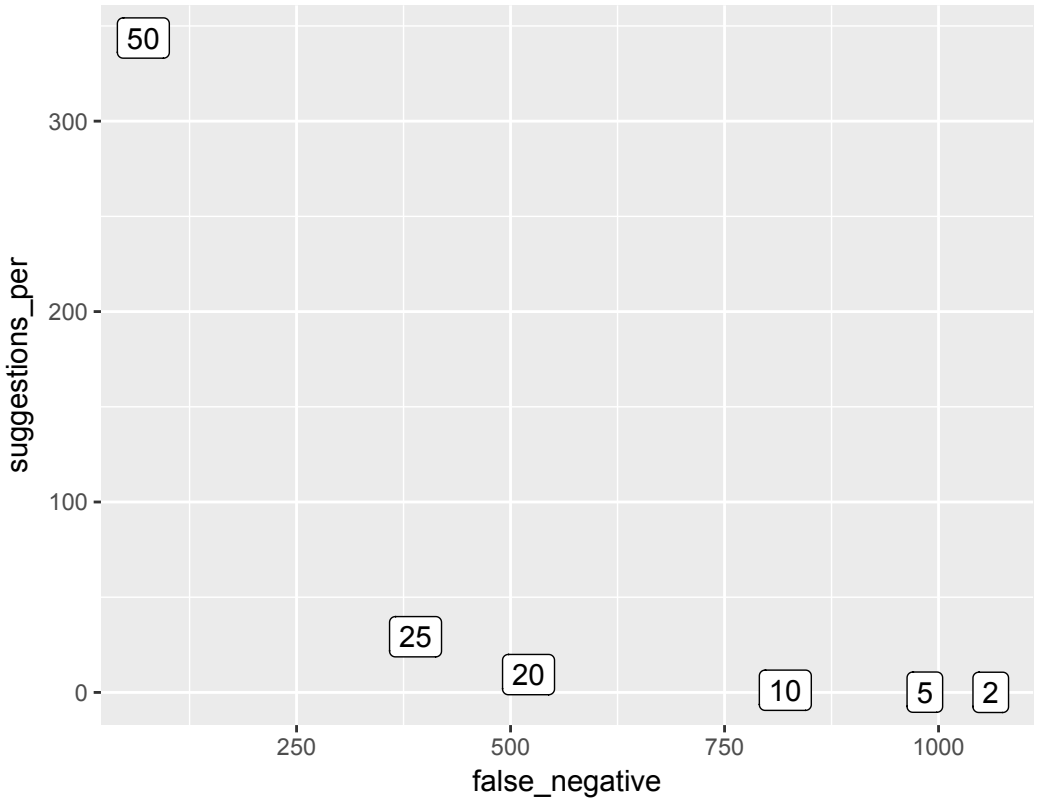

Figure 4: This figure shows the results of a grid search over number of bands used in locality sensitive hashing. The number of minihashes is fixed at 100, and so the number of rows is the number of minihashes divided by the number of bands, which we vary over 50,25,20,10,5 and 2. The outcomes are the number of total suggested matches returned (both true and false) per event location, and the number of false negatives (true matches that are accidentally rejected). Using an elbow rule, the best trade off appears to be 25 bands and 4 rows.

	N	Mismatch	Match
Handlabeled Training Dataset	20,242	17,051	3,192
Exclude Exact Matches	18,405	17,047	1,358
Reverse Diads and Drop Duplicates	34,496	32,374	2,122
Training Split	28,975	27,161	1,814
Test Split	5,521	5,213	308

Table 3: Summary statistics of training and test splits of hand labeled toponym match data

Feature	
First Mismatch	Counts the number of exact matching characters from the beginning of each string
Cosine Distance	The cosine distance (method='cosine') is computed as 1-x\cdot y/(|x||y|), where x and y were defined above.
Levenshtein distance	The Levenshtein distance (method='lv') counts the number of deletions, insertions and substitutions necessary to turn b into a.
Optimal String Alignment distance	Like the Levenshtein distance but also allows transposition of adjacent characters. Here, each substring may be edited only once.
Full Damerau-Levensthein distance	Is like the optimal string alignment distance except that it allows for multiple edits on substrings.
longest common substring	The lcs-distance is defined as the number of unpaired characters. The distance is equivalent to the edit distance allowing only deletions and insertions, each with weight one.
Jaro distance	The Jaro distance is defined as 1-(1/3)(w_1m/|a| + w_2m/|b| + w_3(m-t)/m). Here,|a| indicates the number of characters in a, m is the number of character matches and t the number of transpositions of matching characters.

Table 4: Linker Features. Descriptions taken from the stringdist R package.

Levenshtein, Longest Common Substring, q-grams 1-5, cosine, Jaccard, and a count of the number of matching letters before the first mismatch). We include a count of the number of characters of A and B and the difference between the lengths.

The classifier we employ is a tree based gradient boosting algorithm called Extreme Gradient Boosting (XGBoost) (Chen et al. 2016). The method, in brief, is a greedy function aproximator, additively combining multiple functions, that are estimated one at a time, each seeking to account for the residuals of the prior functions (Friedman 2001). In our case, each function is approximated by a nonparemtric decision tree which starts at a root node containing every observation, and progressively splits observations into purer and purer subsets using cut points along their covariates. We employ a custom loss function, seeking to minimize the log likelihood. We account for the class imbalance by proportionally weighting positive and negative cases. To avoid over-fitting, we use early stopping, based on area under the receiver operating characteristic curve (AUC) on the test set. We observe convergence at about 20 iterations.

The model performs well, with a classification error rate of 0.014. The errors

are balanced across classes, with an F-score of 0.99, an AUC of 0.995, and an area under the Precision Recall Curve of 0.931.

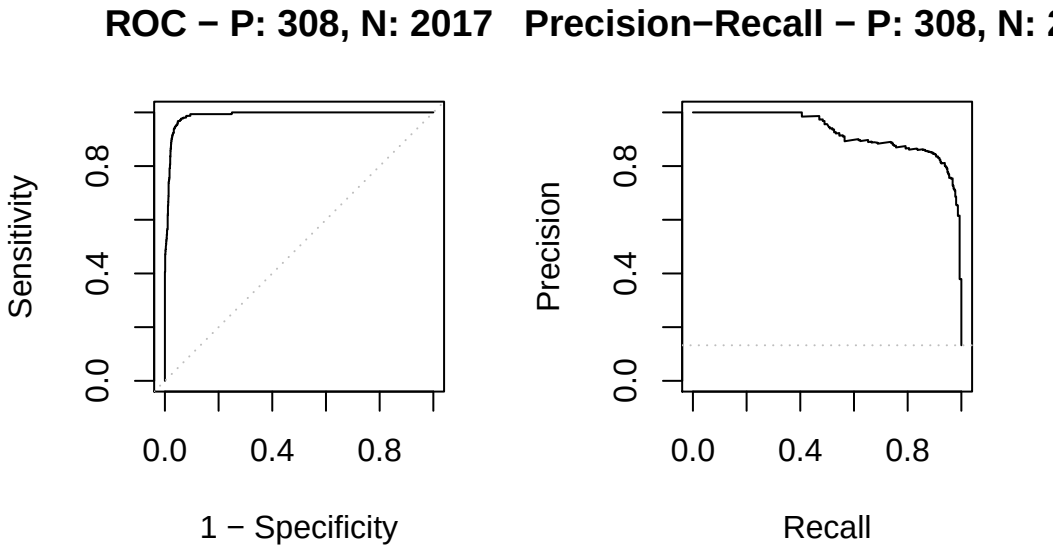

Figure 5: AUC and Precision Recall curves for toponym fuzzy matcher stage 2, predicting likelihood of a match.

Four features contribute disproportionately, Jaro Distance of the stems, the cosine distance of 2-character ngrams of thes stems, 2 and then 3 character ngrams of the full toponym, and then a gradual decline from there on. Through combining different string distance measures, the model is deriving ways in which two strings might match, by both starting with the same word, sharing long substring toward the end, requiring small spelling changes to match, etc.

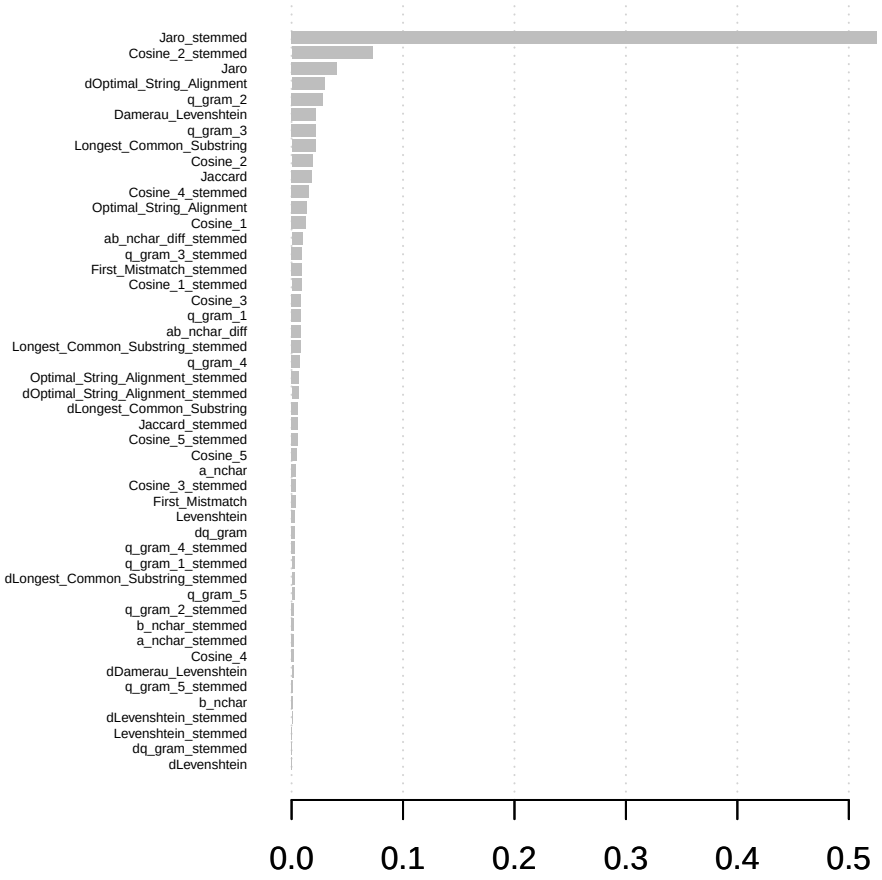

Figure 6: Variable importance measures for toponym fuzzy matcher stage 2, based on XGBoost classifier.

3 "Ensembles"

3.1 "Supervised Ensemble"

We introduce a supervised learning approach to choosing the optimal gazeteer match from among multiple possible matches. The unit of analysis is the location-gazeteer suggestion diad. In total we have 539,268 diads, of which 340,355 are labeled because of the availability of known military coordinate. The outcome is the observed distance between the gazeteer suggestion and the known true location (min 0km, mean 44 km, median 28 km, max 747km). We divide the labeled cases into five mutually exclusive approximately equal folds for cross validation. We split on unique event location names, so that multiple matches to the same event location text or multiple event location texts with different military coordinates don't appear in both the training and validation sets. Therefore the hold out sets aren't exactly balanced, ranging from size 34 thousand to 114 thousand.

The features we employ are: gazeteer source, geometry type, fuzzy or exact string match, and self reference; reporting district, the year of the event, and what level geographic aggregation the event was reported at, e.g. city, district, or province level. We further include the predicted probability of a match genderate by the model from the second stage of the fuzzy toponym matcher. Additionally, we include a one-hot encoding of the gazeteer feature type (PPL=54,391, STM=31,207, ADMD=19,009, etc.).

The model we employ is again XGBoost, implemented in the R package xgboost (Chen et al. 2016). We minimize the mean squared error (on the log transformed distance in kilometers). To avoid over-fitting, we employ early stopping after 10 rounds, converging in 63 iterations. The model has an out of sample 5 fold cross-validated root mean squared error of 1.033 log km. Its predictions are monotonic but not linear in the true distance, shown in the figure below. We therefore find it plausible that the ensemble will do a good job in rank ordering matches from best to worst, but perhaps not guessing correctly how far away a particular match is. The supervised ensemble's rank order correlation with true distance is 0.47, compared to the set of fixed handmade rules whose rank order correlation is 0.37. Visual inspection of the predictions and residuals on a case by case basis reveal the model does a good job distinguishing top ranking matches from poor alternatives, but there is little information available to further sort between poor matches. Poor matches tend to be indistinguishably bad from one another.

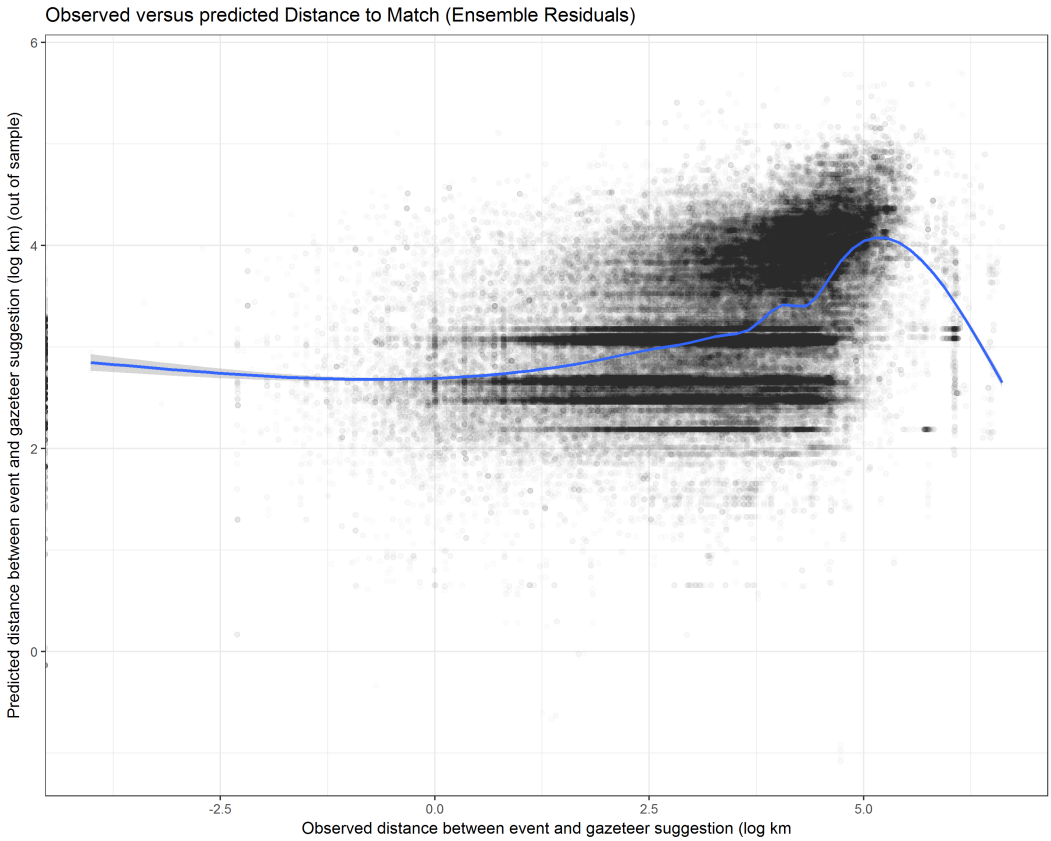

Figure 7: Predicted distance between event and gazeteer match versus observed true distance.

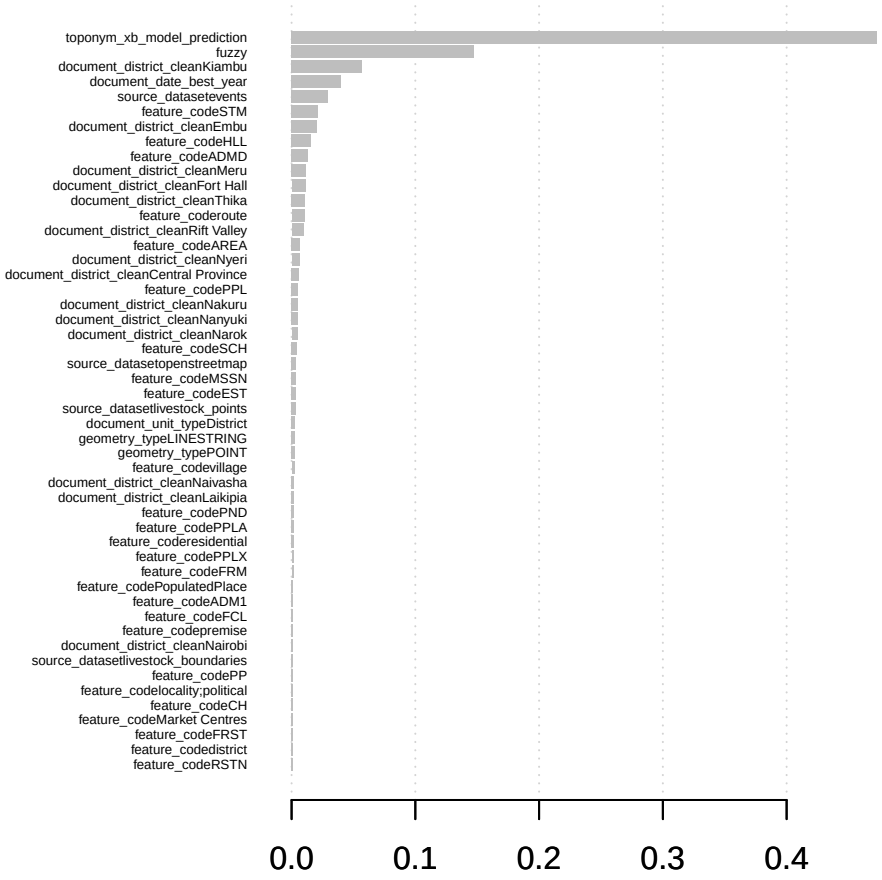

Figure 8: Variable importance measures for supervised ensemble match selection. Categorical variables one-hot encoded. Top 50 shown.

4 "New data on the Mau Mau: The data generating process"

Here we descibe the ontology we developed for coding natural language text descriptions of actions into a consistent series of distinct events. This is for illustration only. Please see the online R-Notebooks for the precise cleaning steps and conversions that were employed.

4.1 Event Types

- Physical Violence
  - Abduction • document language: "abduction", "kidnapping" • description: individuals, regardless of identity or participation on either side of the conflict, were forcefully abducted by participants in the conflict
  - Assault • document language: "assault", "attack" • description: individuals, regardless of identity or participation on either side of the conflict, were physically assaulted by participants in the conflict
  - Murder • document language: "murder", "elimination" • description: individuals, regardless of identity or participation on either side of the conflict, were killed by participants in the conflict
- Property Destruction
  - Arson • document language: "arson", "burn" • description: suspected participants in the conflict set fire to buildings
  - Cattle Slashing • document language: "cattle slashing", "slashed" • description: suspected participants in the conflict killed cattle by slashing them, usually with a panga (similar to a machete)
  - Vandalism • document language: "vandalism" • description: suspected participants in the conflict vandalized private or public property
- Property Theft • document language: "theft" • description: suspected participants in the conflict stole public or private property
- Rebel Captures • document language: "capture" • description: suspected rebels were captured by government authorities, security forces, or civilians
- Oathing • document language: "oathing", "oath". "recruitment" • description: Mau Mau insurgents routinely entered communities and administered "oaths," sometimes voluntarily and sometimes through coercion. Oathing was a traditional practice in Kikuyu, Embu, and Meru communities that bound the participants to their oaths through the threat of

religious and spiritual punishment. The Mau Mau administered two types of oaths: the first was administered broadly to civilians and bound them to silence. Reporting rebel movements, plans, or activities to government officials was considered a violation of the oath. The second level of oath was only administered after the first oath and only to those committed to participating in Mau Mau. It was referred to as the Batuni (or killing) oath. The intelligence reports do not routinely distinguish between the two types of oaths.

- Security Operations
  - • Contact • document language: "contact", "drove off", "chased off", "ambush" • description: security forces, usually on patrol or as a part of operations, made contact with suspected rebel forces. Sometimes this resulted in a violent engagement and other times the rebel forces simply fled. [***keep an eye out to make sure "drove off" is being categorized correctly***]
  - • Patrols • document language: "patrol", "sweep" • description: security forces sent units on patrol regardless of whether they made contact. If contact was made, the patrol was coded as a contact rather than a patrol.
  - • Punishment • document language: "confiscate", "sentenced" • description: government authorities or security forces confiscated property (presumably of suspected insurgents or their sympathizers) or suspected insurgents received punishments through the colonial legal system (although this was usually not reported in the intelligence documents)
  - • Screening • document language: "screening" • description: Colonial authorities set up screening panels in local communities and interviewed individuals to try and determine whether they were Mau Mau or rebel sympathizers. Those deemed "guilty" by the screening committee were often sent to detention camps or tried and sentenced through the legal system.
- Unclassified (at mid-level of aggregation)
  - Desertion • document language: "desertion" • description: soldiers deserting their units
  - Escape • document language: "escape" • description: suspected rebels escaping from civil authorities, security units, or others attempting to capture them

4.2 Instigator/Target Types

- Government
  - Civil Authorities

    * Colonial Authorities • document language: “councillor”, “district commissioner”, “district officer” or “d.o.” or “do”, “forest ranger”, “game ranger”, “game warden”, “government”, “government employees”, “port authority”, “public works department”, “screening team”, “wakamba screening team” • description: employees and officials of the colonial government excluding tribal authorities and members of the security forces.
    * Tribal Authorities • document language: “chief”, “elders”, “headman” • description: chiefs, headman, and their personal details were part of the tribal authority structure that worked with the colonial government to govern the various tribal areas. Many were loyal to the colonial government during the emergency while others were suspected or accused of clandestinely supporting the Mau Mau.
    * Home Guard • document language: “Embu guard”, “farm guard”, “forest guard”, “home guard”, “Ikandine guard”, “Kathanjure guard”, “Kijabe guard, “Kikuyu guard”, “Masai guard”, “Meru guard”, “Nandi guard”, “Nkubu guard,” “stock guard”, “Tigoni guard” (and abbreviations such as e.g., h.g., k.g., m/g, ng, etc.) • description: home guards were local defense forces recruited and based at the village or sublocation administrative level and were usually headed by the local chief or headman. Many were referred to by the ethnic group which they were recruited from (Embu, Kikuyu, Masai, Meru, Nandi, etc.). Others were referred to by their primary duty (farm guard, forest guard, etc.) or by their location (Kathanjure guard,etc.).

– Military
    * Arab combat units • document language: “arab combat” • description: military units recruited from Middle Eastern immigrants and their descendants
    * Asian combat units • document language: “asian combat”, “asian combat team”, “asian combat unit” • description: military units recruited from Indian and Pakistani immigrants and their descendants
    * British Military • document language: “Devonshire regiment” or “devons” or “ a co devon”, “Field Intelligence Assistant” or “fia”, “Field Intelligence Officer” or “FIO” or “fios”, “Gloucestershire regiment” or “glosters”, “Lancashire fusiliers”, “King’s Shropshire light infantry” or “ksli”, “Royal East Kent regiment” or “buffs”, “Royal fusiliers”, “Royal Highland regiment” or “Black Watch” or “watch” or “bw”, “Royal Inniskilling fusiliers” or “royal innisks” or “inniskillings”, “Royal Irish fusiliers”, “Royal Northumberland fusiliers” or “RNF” (and abbreviations such as BW, LF, RIF, etc.), “army”, “platoon support company”, “c company”, “d company”, “a company”, “d force” • description: units of the

British Army whose members were recruited in the United Kingdom and were sent to Kenya to support locally recruited forces. For a complete listing of such units see http://www.britains-smallwars.com/kenya/Units.html.

  * King's African Rifles • document language: 3 KAR, 4 KAR, 5 KAR, 6 KAR, 7 KAR, 23 KAR, 26 KAR (and variations) • description: The King's African Rifles were East African colonial military units. While they were officered almost exclusively by British officers, their rank-and-file were recruited locally from African communities in Kenya, Uganda, and Tanganyika. The 3rd, 5th, 7th, 11th, and 23rd battalions were recruited from Kenya, the 4th from Uganda, and the 6th and 26th from Tanganyika. Kenya Regiment • document language: "Kenya regiment", "kenreg", "kenregg", kenya regiment sergeant", "kr", "Captain Folliott's team" • description: the Kenya Regiment was a military unit recruited from local settlers of European descent. Conscription began in 1950 from the European settler community. Officers were often seconded to the KAR battalions.
  * Military (generic) • document language: "captain", "company" or "coy", "military", "military property", "platoon" (and abbreviations), "security forces", "security force", "sentry", "striking force" • description: military units that could not be coded in a more specific category because the document language was too vague
  * Pseudo Gangs • document language: "pseudo gang", "pseudo gangs", "pseudo team", "pseudo teams", "tracker group," "trojan", "trojan team" • description: the British military developed special forces called "pseudo gangs" that went into the forest disguised as Mau Mau gangs in order to gather intelligence and infiltrate actual Mau Mau gangs.
  * Royal Air Force • document language: "RAF", "raf lincolns", "bombers", "air strike", "flying squard", "harvards" (a type of aircraft) • description: British RAF units
  * Paramilitary • document language: "General Service Unit", "GSU" • description: the General Service Unit was a special forces paramilitary unit recruited from both police and military units. It was comprised of white officers and African rank-and-file.

- Police
  * CID • document language: "CID" or "cid" • description: Criminal Investigations Division of the police
  * Kenya Police • document language: "Kenya police", "KP" • description: The regular Kenya police were a mixed force of white settlers and Africans.
  * Kenya Police Reserve • document language: "Kenya police reserve", "KPR", "KPR officers", "reserve police officer" or "RPO"

or "rpos" • description: The Kenya Police Reserve was formed for emergency purposes and was recruited from the European settler committee. Settlers could fulfill their conscription duties by serving in the KPR.

  * Police (generic) • document language: "constable", "police", "police party" • description: police units that could not be coded in a more specific category because the document language was too vague
  * Railway Police • document language: "railway police" • description: special police units for the railway
  * Special Branch • document language: "sb officers", "special branch", "special branch team", "blue doctor team" • description: Special Branch was an intelligence unit within the colonial police forces responsible for gathering information on subversion and other threats to the state. They conducted interrogations, analyzed captured documents, and compiled and distributed regular intelligence reports and appreciations from both police and military intelligence sources.
  * Tribal Police • document language: "African special constable", "Githumu police", "Masai special constable", "tribal police", "TP", "tpeg" • description: tribal police were locally recruited units of Africans, organized at the district level, that helped maintain law and order in the native reserve areas. African constables from the Kenya Police were often seconded to tribal police units as commanders. The tribal police officially came under the command of the District Officer or District Commissioner.
  * Tribal Police Reserve • document language: "tribal police reserve", "TPR" • description: additional tribal police units that supported the tribal police during the emergency. The were recruited from the local African population.

- Civilians
  * Civilians • document language: "african", "Africans", "children", "civilians", "driver", "embu", "employees", "european", "evangelist", "family", "farm boys", "farm labour", "farmer", "girls", "herd boys", "houseboy", "informer", "Kikuyu", "kuria tribesman", "labour", "laborour", "local labour", "loyalist", "manager", "Masai", "masai party", "men", "mission staff", "mrhiggins", "owner", "passengers", "people", "shopkeeper", "samburu", "sikh", "stranger", "students", "teachers", "tribesmen," "Turkana", "vet officer", "vigilantes", "woman", "women", "workers" • description: individuals identified in the documents specifically as civilians or non-participants in the conflict.
  * Communities • document language: "camp", "manyatta", "fishing camp", "sublocation", "village" • description: sets of individuals identified in the documents by a communal affiliation.

- Suspected Rebels
  - Detainees • document language: “detainees” • description: individuals detained by the government on suspicion of insurgent activity.
  - Suspected Insurgents • document language: “bandits”, “food foragers”, “gangs”, “gunmen”, “Kiama Kia Muingi” or “KKM” (offshoot of Mau Mau later in the conflict), “komerera” (means ill disciplined fighter), “Mau Mau”, “oath administrator”, “passive wing”, “passive wing members”, “rebels”, “resistance groups”, “suspects”, “terrorist”, “terrorists” • description: individuals identified in the documents as insurgents. The colonial authorities and security forces were quick to identify as an insurgent any individual they harmed or who they found committing a crime in a conflict area or in possession of a weapon (including pangas which were used for harvesting). Thus this category likely includes many civilians as well as actual insurgents.
- Property
  - Armaments
    - Ammunition • document language: “ammunition” • description: bullets and other ammunition for firearms.
    - Explosives • document language: “explosives”, “gelignite” • description: materials used for building explosives.
    - Firearms • document language: “arms”, “firearm”, “gun”, “pistol”, “rifle” • description: any type of firearm.
    - Other Weapons • document language: “axe”, “panga”, “scabbard”, “weapons” • description: any other type of weapon not considered a firearm or an explosive.
  - Private Property • document language: “buildings”, “cattle boma”, “cattle dip”, “coffee trees”, “duka”, “farms”, “garage”, “homes” or “huts”, “hotel”, “land rover”, “lorry”, “market”, “office”, “oxcart”, “property”, “pump house”, “sawmill”, “shops” or “stores”, “tractor”, “vehicle”, “windmill” • description: any property that could be considered private (excluding government and military property, infrastructure such as roads, and property used for public use even if privately owned such as churches and schools).
  - Provisions
    - Cash • document language: “cash”, “funds”, “money” • description: currency.
    - Food • document language: “banana”, “barley”, “bran”, “cabbage”, “coffee”, “corn”, “cream”, “crops”, “dairy”, “food”, “fruit”, “grain”, “honey”, “maize”, “meat”, “milk”, “oats”, “posho”, “potatoes”, “sugar”, “vegetable”, “wheat” • description: food items.
    - Livestock • document language: “beast”, “cattle”, “cow”, “herd”, “livestock”, “pig”, “sheep”, “steer”, “stock” • description: livestock.

    - Medicine • document language: “medical supplies”, “medicine”, “m&b tablets” • description: pharmaceuticals and other medical supplies.
    - Supplies • document language: “bags”, “bedding”, “blankets”, “books”, “charcoal”, “cloth”, “clothing”, “cooking utensils”, “cutlery”, “dairy item”, “equipment”, “farm implements”, “household items”, “instruments”, “iron”, “pails”, “petrol”, “provisions”, “oil”, “railway uniforms”, “sacks”, “supplies”, “tarpaulin”, “thatch,” “timber”, “tobacco”, “tools”, “uniforms”, “wire”, “wireless set”, “whiskey” • description: general household and farm goods.
- Public Buildings
    - Church • document language: “church” • description: churches.
    - Infrastructure • document language: “airstrip”, “bridges”, “half built village”, “military property”, “prison camp”, “roads”, “trenches”, “water tank” • description: general government or security forces infrastructure.
    - School • document language: “school” • description: schools (including both private and religious schools).

5 Evaluating the Consequences of Georeferencing Decisions

5.1 “Predictability of Missingness”

In the paper we present results from a model predicting the likelihood of an event to be missing exact military coordinates as a function of a number of properties of the event as well as the document it was drawn from. We show that location information is not missing at random, it is missing systematically, and systematically in ways that are correlated with both outcomes, covariates, and confounders that would go in any typical downstream analysis of these data. We replicated this excercise for other types of missigness (military coordinates, text labels, coordinates and labels), and for missingness following imputation by every strategy and decsision discussed. Those results are presented in the figure below. Again the model is XGBoost implemented in the R package xgboost (Chen et al. 2016). The unit of analysis is the event. The loss function is log loss with proportionally weighted cases to account for class imbalance. The measure of accuracy of the model (predictability of missingness) is Area Under the Receiver Operating Characteristic Curve. Values near 0.5 are close to random accuracy, and values close to 1 are perfect accuracy.

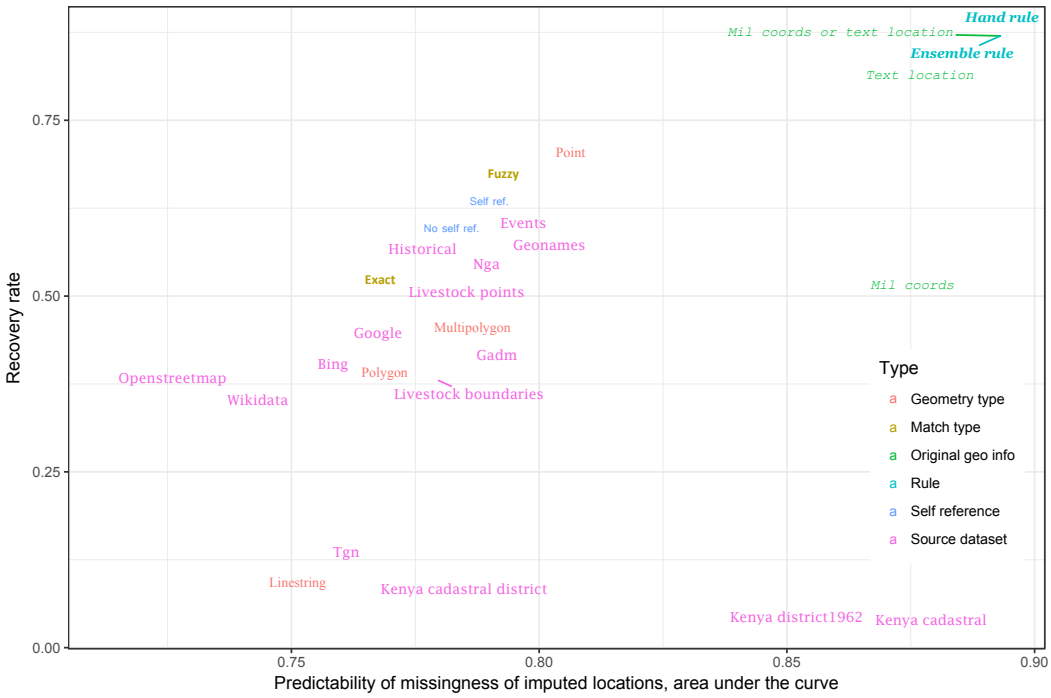

Figure 9: This figure shows the predictability of missingness of location information for the raw data as well as each imputation strategy. The features are properties of each event, such as the perpetrator, target, type of action etc. and properties of the document from which it was drawn, like reporting district, year, time period covered, etc.

References

[1] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. ACM, 2016.

[2] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive datasets. Cambridge university press, 2014.

[3] Santos, Rui, Patricia Murrieta-Flores, and Bruno Martins. "Learning to combine multiple string similarity metrics for effective toponym matching." International Journal of Digital Earth (2017): 1-26.