

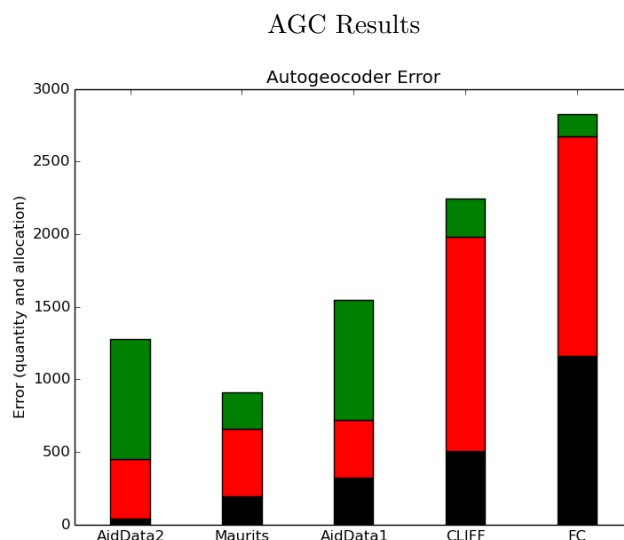
AutoGeocoding Update

This document briefly summarizes the findings from our initial assessment of the autogeocoding activity. Data limitations significantly hindered our capacity to conduct these comparisons, and thus limit the conclusions of this report¹. Noting these limitations, we draw three conclusions and make two recommendations.

1. **Data Providence and Chains** - previous human geocoding efforts did not provide enough information to link project documents to individual geocoded locations. This was due to a lack of specificity in the linkages ('See Google Drive'), a failure to save the information used to generate geocodes, and a lack of clear standards for geocoders. Further, there is no consistent record of inter- (or intra-) coder reliability amongst our human coders. This severely hampers efforts to determine the relative accuracy of geocoding, or even to test autogeocoders, and limits the generalizeability of this report.
2. **Scale** - there are many scales at which geocoders may be relevant - from prediction of individual points, to various administrative or other levels. We found that finer degrees of granularity are increasingly hard for geocoders to ascertain; however, estimates of the total quantity of projects in an area were found to be highly accurate. At different scales, autogeocoding could play different roles in AidData workflows. At the lowest level, it appears it can provide very good estimates for the total LOE a project may take for human coders, given a set of input documents, by producing an estimate of the number of projects contained in that documentation. As higher levels of spatial accuracy are needed, the capacity of autogeocoders appears to degrade, but the true rate of this degradation is currently unknown due to data providence concerns.
3. **Existing GeoCoders** - in contrasting multiple geocoders, it was found that the coder commissioned from Maurits was most succesful in this set of tests in terms of total error at the finest administrative scale tested (Malawi / ADM2). However, it did not perform best in other aspects (ADM1 and total quantity error) when contrasted to other, off-the-shelf or otherwise relatively easy-to-implement models.

Recommendations

1. **Data Providence and Chains** - We highly recommend that all data used to produce geocoded information is retained, and individual projects are linked to the relevant input information. Some activities are already ongoing to support this - the Data Team is working closely with student geocoder managers to ensure the full chain of the data product is saved. Further, onlining of disk space by VPR will allow for a permanent repository for this data. Once these datasets are available, the REU will work with the Data Team and Geocoders to extract a sample of source documents from projects geocoded during summer 2015. This sample will be used for future tests to allow for absolute comparisons of accuracy, as well as enhance the external validity of these comparisons.
2. **GeoCoding Funding** - The relatively strong performance of Maurits' code suggests that gains do remain to be had by continuing to pursue more specialized autogeocoding approaches, if finer scale information is desired. However, the true rate of gain (in terms of percentage accuracy) is currently unknown due to the limitations of the assessment data. Further, initial efforts by in-house researchers suggest that large gains may be achievable with relatively low degrees of effort. Until better assessment data can be acquired (ETA summer 2015), it is recommended that we do not continue to fund external GeoCoding efforts, and maintain only a minor in-house focus to prevent a loss of institutional knowledge. After the summer 2015 information is acquired, it will become feasible to re-engage (potentially in the form of a "grand challenge"), as we will have a more reliable dataset against which to test.



Black bars represent quantity error; red bars additional error introduced by allocating points to ADM1; green bars additional error at ADM2. The total height of the bar represents total error at ADM2.

¹These results are built upon a subset of all documents used to geocode Malawi, contrasted to the final produced Malawi dataset. As such, it was impossible for any geocoder to attain perfect accuracy, and it is unknown what the delta is between achievable accuracy and perfect accuracy. These results can only be interpreted as (a) a worst-case example for geocoding, and (b) a relative assessment of the capacity of various geocoders. They should not be interpreted as an estimate of the absolute accuracy of any particular geocoder.